Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

Wiebke Manuela Schulze, M. Sc.

born in Braunschweig

Oral examination: 13.09.2017

MUTUALLY EXCLUSIVE CBC AND CBC-ARS2 CONTAINING COMPLEXES COORDINATE THE FATE OF RNA POLYMERASE II TRANSCRIPTS

Referees:

Prof. Dr. Iain Mattaj

Prof. Dr. Georg Stoecklin

ABSTRACT

RNA Polymerase II (Pol II) transcripts comprise many different RNA classes that undergo co-transcriptionally complex maturation processes including the addition of a 5′ cap structure, removal of introns via splicing, 3′ end processing and the assembly into transport competent ribonucleoprotein particles. How all these processes are coordinated, regulated and quality controlled, is a central question in RNA biology. One of the processes all Pol II transcripts have in common is the addition of a cap structure to the 5′ end. This 5′ cap structure is immediately bound by the heterodimeric cap binding complex CBC comprising subunit CBP80 and the cap binding subunit CBP20. The cap structure as well as CBC are crucial for the correct processing of the nascent transcript. The subsequent maturation processes including the involved processing machineries depend on the RNA class. However, some proteins like NELF-E (negative elongation factor subunit E) and ARS2 (arsenite-resistance protein 2) seem to associate with the RNA-CBC complex regardless of the type of RNA, whereas other CBC interacting proteins like the phosphorylated adaptor for RNA export PHAX are important for distinct RNAs.

To better understand the individual CBC complexes, different biochemical, biophysical and structural methods were used to analyse the individual interactions between CBC and ARS2, PHAX as well as NELF-E on molecular level. The studies showed the incompatibility of ARS2 and NELF-E binding to CBC and revealed that both proteins are sharing the same binding site on CBC and that their interacting residues show high conservation and similarities. In addition, solving the structure of human ARS2 allowed the identification of domains and residues involved in RNA and protein-protein interactions. Further analysis also led to the identification of a novel CBC-ARS2 complex, including the nuclear cap binding protein NCBP3, namely CBC-ARS2-NCBP3. The results obtained support the hypothesis that ARS2 serves as additional binding surface for proper sorting and processing of nascent transcripts. Additionally, all results together enable to propose a model of RNA processing containing mutually exclusive RNA-CBC complexes based on the RNA maturation stage as well as the RNA class.

ZUSAMMENFASSUNG

Viele verschiedene RNA Klassen werden von der RNA Polymerase II (Pol II) synthetisiert. All diese Pol II Transkripte durchlaufen co-transkriptional zahlreiche komplexe Reifungsprozesse wie die Modifikation des 5′ Endes durch Anfügen einer Cap Struktur, das Entfernen von Introns durch Spleißen, die 3′ Ende Modifizierung und die Bildung von Transport-kompetenten Ribonukleoprotein-Partikeln. Wie all diese Prozesse koordiniert, reguliert und kontrolliert werden ist eine zentrale Frage innerhalb der RNA Biologie. Eine Weiterverarbeitung haben alle Pol II Transkripte gemeinsam: die Addition der Cap Struktur am 5′ Ende. Diese 5′ Cap Struktur wird sofort von dem Cap bindenden Komplex CBC gebunden, welcher aus CBP80 und CBP20 besteht. Die Cap Struktur genauso wie CBC ist essentiell für die korrekte Weiterverarbeitung des entstehenden Transkriptes. Die weiteren Reifungsprozesse und die damit verbundenen Prozessierungskomplexe sind abhängig von der RNA Klasse. Allerdings scheinen einige Proteine wie NELF-E (Untereinheit E von NELF (negative elongation factor)) und ARS2 (arsenite-resistance protein 2) den RNA-CBC Komplex unabhängig von der RNA Art zu binden, während andere Proteine wie der snRNA Export Adapter PHAX spezifisch für individuelle RNA Klassen sind.

Um die individuellen CBC Komplexe besser zu verstehen, wurden die Interaktionen zwischen CBC und ARS2, PHAX und NELF-E durch verschiedene biochemische, biophysikalische und strukturbiologische Methoden auf molekularem Level untersucht. Die Studien zeigten, dass die Interaktion von ARS2 und NELF-E mit CBC nicht kompatibel ist und dass dies durch eine gemeinsame Bindungsstelle von NELF-E und ARS2 an CBC verursacht wird. Zudem weisen die CBC-ARS2 und CBC-NELF-E Interaktionen viele Gemeinsamkeiten auf. Die mittels Kristallographie determinierte Struktur des humanen ARS2 diente der Identifikation von Domänen und Aminosäureresten, die in RNA-Protein und Protein-Protein Interaktionen beteiligt sind. Dies führte unter anderem zu der Entdeckung eines neuen CBC-ARS2 Komplexes, dem CBC-ARS2-NCBP3 Komplex. Des Weiteren unterstützen all diese Ergebnisse ein Model der RNA Weiterverarbeitung, in dem sich, abhängig von dem Reifegrad der RNA und der RNA Klasse, gegenseitig ausschließende RNA-CBC Komplexe bilden.

PUBLICATIONS

Some results and figures have appeared previously and are about to be published in the following publications:

[1] S. Giacometti, N. E. Benbahouche, M. Domanski, M. C. Robert, N. Meola, M. Lubas, J. Bukenborg, J. S. Andersen, W. M. Schulze, C. Verheggen, G. Kudla, T. H. Jensen, and E. Bertrand. "Mutually Exclusive CBC-Containing Complexes Contribute to RNA Fate". In: *Cell Reports* 18.11 (2017), pp. 2635–2650.

[2] W. M. Schulze and S. Cusack. "Structural basis for mutually exclusive, co-transcriptional nuclear cap-binding complexes with either NELF-E or ARS2". In submission to Nature Communcations. 2017.

[3] W. M. Schulze and S. Cusack. "Structural analysis of ARS2". In preparation.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACN	acetonitril
Amp	ampicillin
AP	affinity purification
ARS2	arsenite-resistance protein 2
ATP	adenosine triphosphate
BrdU	bromodeoxyuridine
BSA	bovine serum albumin
CAPS	3-(cyclohexylamino)-1-propanesulfonic acid
CBC	cap binding complex
CBP	cap binding protein
CDK	cyclin-dependent kinase
CE	capping enzyme
CFI_m	cleavage factor I_m
CIP	calf intestinal phosphatase
CK2	casein kinase 2
CPSF	cleavage and polyadenylation specificity factor
CRM1	chromosomal region maintenance 1
CstF	cleavage stimulatory factor
CTD	C-terminal repeat domain
CTR	C-terminal region
DAPI	4′,6-diamidino-2-phenylindole
DMEM	Dulbecco's Modified Eagle Medium
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DRB	dichloro-1-β-D-ribofuranosylbenzimidazole
ds	double stranded
DSE	distal sequence element
DSIF	DRB sensitivity-inducing factor
DSS	disuccinimidyl suberate
DTT	dithiothreitol
EDTA	ethylenediamine tetraacetic acid
EGFP	enhanced GFP

eIF4G	eukaryotic translation initiation factor 4G
EJC	exon junction complex
ESRF	European Synchrotron Radiation Facility
FA	formic acid
FADD	Fas associated protein with death domain
FAM	fluorescein amidite
FARB	FLASH-ARS2-binding
FBS	fetal bovine serum
FLASH	FLICE-associated huge protein
FLICE	FADD-like IL-1β-converting enzyme
GFP	green fluorescent protein
GMP	guanosine monophosphate
GST	glutathione S-transferase
GTase	guanylyltransferase
GTFs	general transcription factors
HDE	histone downstream element
HEK	human embryonic kidney
HEPES	4-(2-hydroxyethyl)piperazine-1-ethanesulfonic acid
hnRNP	heterogeneous nuclear ribonucleoprotein
HTX	high-throughput crystallisation
IPTG	isopropyl-β-D-thiogalactopyranosid
ITC	isothermal titration calorimetry
Kan	kanamycin
LB	lysogeny broth
ld	linear discriminant
lincRNA	long intergenic/intervening non-coding RNA
lncRNA	long non-coding RNA
m^7G	7-methylguanosine
MALS	multiangle light scattering
MBP	maltose binding protein
MIF4G	middle domain of eIF4G
miRNA	microRNA
MOPS	3-(N-morpholino) propanesulfonic acid
MPD	2-methyl-2,4-pentanediol
mRNA	messenger RNA

mRNP	messenger ribonucleoprotein
N7MTase	guanine-7-methyltransferase
NBE	NELF-E binding element
NCBP3	nuclear cap binding protein 3
NELF	negative elongation factor
NES	nuclear export signal
NEXT	nuclear exosome targeting
NLS	nuclear localisation signal
NMR	nuclear magnetic resonance
NPC	nuclear pore complex
nts	nucleotides
OD	optical density
P-TEFb	positive transcription elongation factor b
PABP	poly(A) binding protein
PAGE	polyacrylamide gel electrophoresis
PAP	poly(A) polymerase
PAS	polyadenylation signal
PAXT	poly(A) tail exosome targeting
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PDB	protein data bank
pds	partially double stranded
PEG	polyethylene glycol
PEI	polyethylenimine
PFA	paraformaldehyde
PHAX	phosphorylated adaptor for RNA export
PIC	preinitiation complex
piRNA	PIWI-interacting RNA
PIWI	P-element-induced wimpy testis
Pol II	RNA polymerase II
PP2A	protein phosphatase 2A
PROMPTs	promoter upstream transcripts
PSE	proximal sequence element
pre-mRNA	precursor messenger RNA
pre-snRNA	precursor small nuclear RNA

pri-miRNA	primary microRNA
RBD	RNA binding domain
RDH	replication-dependent histone
RMSD	root-mean-square deviation
RNA	ribonucleic acid
RNP	ribonucleoprotein
RPAP2	RNA polymerase II associated protein 2
RRM	RNA recognition motif
rRNA	ribosomal RNA
RT	room temperature
RTPase	RNA 5′-triphosphatase
SAH	S-adenosylhomocysteine
SAM	S-adenosylmethionine
SAXS	small angle X-ray scattering
SDS	sodium dodecyl sulfate
SEC	size exclusion chromatography
SIR	single isomorphous replacement
SLBP	stem-loop binding protein
SNAPC	snRNA activation protein complex
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
SRP	signal recognition particle
ss	single stranded
SUMO	small ubiquitin-like modifier
TCEP	tris(2-chloroethyl) phosphate
TEV	tobacco etch virus
TF	transcription factor
TFA	trifluoroacetic acid
TMT	tandem mass tag
TRAMP	Trf4/Air2/MTR4 polyadenylation
TREX	transcription-export
tRNA	transfer RNA
tss	transcription start site
UTR	untranslated region
UV	ultraviolet

AMINO ACID ABBREVIATIONS

For amino acids the three letter and the one letter code was used:
A, Ala: alanine
C, Cys: cysteine
D, Asp: aspartate
E, Glu: glutamate
F, Phe: phenylalanine
G, Gly: glycine
H, His: histidine
I, Ile: isoleucine
K, Lys: lysine
L, Leu: leucine
M, Met: methionine
N, Asn: asparagine
P, Pro: proline
Q, Gln: glutamine
R, Arg: arginine
S, Ser: serine
T, Thr: threonine
V, Val: valine
W, Trp: tryptophan
Y, Tyr: tyrosine

NUCLEOTIDE ABBREVIATIONS

For the description of ribonucleosides and deoxyribonucleosides, the one letter code was used:
A: adenosine/deoxyadenosine
G: guanosine/deoxyguanosine
C: cytidine/deoxycytidine
T: thymidine
U: uridine
R: purine nucleoside/purine-deoxynucleoside
Y: pyrimidine nucleoside/deoxynucleoside
H: any nucleoside/deoxynucleoside but guanosine
N: any nucleoside/deoxynucleoside

1 INTRODUCTION

1.1 GENE EXPRESSION

Genetic information is encoded in deoxyribonucleic acid (DNA) and the process of transferring this information into functional molecules such as ribonucleic acids (RNAs) and proteins is called gene expression. The initial step of gene expression is the transcription of the genetic information from DNA to RNA. In humans most of the genome is transcribed [42, 75], although protein-coding genes comprise less than 3% of the human genome [36]. Co-transcriptionally nascent RNAs undergo complex maturation processes including the addition of a 5′ cap structure, removal of introns via splicing, 3′ end processing and assembly into transport competent ribonucleoprotein (RNP) particles. To ensure the correct maturation and thus the correct transfer of information, all these processes need to be coordinated, regulated and quality controlled and how this is achieved for different kinds of RNAs is a central question in RNA biology.

1.2 RNA CLASSES

RNA is a biological macromolecule essential for all known forms of life. For a long time, only three RNA types involved in protein synthesis were known: messenger RNA (mRNA) as the carrier of genetic information, transfer RNA (tRNA) linking mRNA and protein biosynthesis, and ribosomal RNA (rRNA) as part of ribosomes. In most eukaryotic cells, these RNA classes account for 98% of cellular RNA mass [216]. However, RNA plays a much bigger role apart from being essential for protein biosynthesis. For example in some viruses RNA instead of DNA is the carrier of the genetic information and RNA can act as an enzyme, called ribozyme. RNAs together with proteins form RNPs that are involved in different processes and can have enzymatic activity (e.g. snRNAs are components of spliceosomes and rRNAs of ribosomes). Additionally, RNAs are involved in the regulation of many cellular processes like differentiation, cell division, growth, aging and death. All these RNA functions are mediated by distinct RNA classes. As RNAs have a very complex role, dysregulation or defects in certain RNAs have been associated with a number of human diseases.

In metazoan cells, in addition to mRNAs, RNA transcripts include non-coding RNAs such as tRNAs and rRNAs, long non-coding RNAs (lncRNAs), microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), PIWI-interacting RNAs (piRNAs), telomerase RNAs and signal recognition particle (SRP) RNAs. At certain cellular stages like embryogenesis, cell division and germ cell development specific RNA levels can either be up or down regulated. Thus, the abundance of individual RNAs changes depending on the cell cycle as well as the cell type.

In eukaryotes the different RNAs are synthesised by three different RNA polymerases, namely RNA polymerase I (Pol I), RNA polymerase II (Pol II) and RNA polymerase III (Pol III). The main transcripts of the different polymerases are the following: 18S, 5.8S and 28S rRNAs are synthesised by Pol I, mRNAs together with snRNAs are synthesised by Pol II while 5S rRNAs and tRNAs are synthesised by Po III.

1.3 RNA POLYMERASE II TRANSCRIPTS

1.3.1 *Role of RNA Polymerase II C-terminal Domain in Transcription Regulation and Co-transcriptional Processing*

Pol II transcripts comprise the precursors of mRNAs and many non-coding RNAs, e.g. snRNAs, miRNAs and piRNAs. Transcription by Pol II is divided into initiation, elongation and termination phases (Figure 1.1) and each stage is linked to RNA processing and regulatory events [88].

Most eukaryotic RNA precursors undergo extensive co-transcriptional maturational processing, including 5′ end capping, splicing, and 3′ end cleavage and polyadenylation. Many nuclear factors involved in these co-transcriptional events are recruited via the C-terminal repeat domain (CTD) of Pol II, which is a unique feature of this polymerase. It is made up of 15 to 52 heptad repeats with the consensus sequence Tyr1–Ser2–Pro3–Thr4–Ser5–Pro6–Ser7 [37] and is subject to hyperphosphorylation at Tyr1, Ser2, Thr4, Ser5 and Ser7 [14, 28, 54, 81, 88, 191]. The phosphorylation state of the CTD changes through the transcription cycle with different patterns of modifications recruiting enzymes and proteins in a transcription-stage specific manner.

At the transcription start site (tss) transcription initiation factors assemble and recruit Pol II, which leads to the formation of the preinitiation complex (PIC) (Figure 1.1). Part of this complex is the cyclin-dependent kinase (CDK)7, the kinase of the transcription factor TFIIH, which phosphorylates the CTD on Ser5 (Ser5P). Ser5P leads to the recruitment of different transcription elongation factors and start of elongation. In addition, the capping enzyme (CE) directly binds via its RNA guanylyltransferase to the Ser5P CTD, which stimulates the activity of the CE [52, 65]. Ser5P is further known to enhance the recruitment of histone lysine methyltransferase Set2, which transfers methyl groups on histone H3 K36 residues [106]. Recently, it was also shown that TFIIH can phosphorylate the CTD on Ser7 (Ser7P) [2, 68, 101]. Ser7P mainly stays in context with the recruitment of the multisubunit Integrator complex for mammalian snRNA 3′ end processing [11, 48].

During elongation, the level of Ser5P decreases while the phosphorylation on Ser2 (Ser2P) increases until Pol II reaches the 3´ end of the transcription unit. Rtr1 [149] and Ssu72 [107] as well as Fcp1 [26] were identified as Ser dephosphatases, with Rtr1 dephosphorylating Ser5P and possibly playing a role in the early elongation phase. Ssu72 was shown to dephosphorylate the remaining Ser5P at the cleavage and polyadenylation sites [107, 215, 225, 232] and Fcp1 to remove the Ser2P when Pol II reaches the 3´ end of genes [35]. Besides, CDK9, subunit of the positive transcription elongation factor b (P-TEFb), was described to phosphorylate Ser2 as well

as the DRB sensitivity-inducing factor (DSIF) [166]. DSIF consists of subunits Spt4 and Spt5, with Spt5 also containing a C-terminal region (CTR) with repeats that can be phosphorylated at Thr4 [226]. The recruitment and kinase activity of P-TEFb are important to promote elongation by releasing the promoter-proximal paused Pol II (see Sec. 1.3.2).

Altogether, the dynamic and stage-specific phosphorylation/dephosphorylation of the Pol II CTD is critical for proper RNA biogenesis.

1.3.2 *Promoter Proximal Pausing of RNA Polymerase II*

One major rate-limiting step within transcription and gene expression is the promoter proximal pausing of Pol II (Figure 1.1). At more than 70% of metazoan genes Pol II pauses after synthesising 20-40 nucleotides (nts) nascent RNA [125]. This process is controlled by the phosphorylation pattern of the CTD of Pol II (see Sec. 1.3.1). Paused Pol II shows an enrichment of Ser5P, whereas enrichment of Ser2P stimulates elongation. DSIF and the negative elongation factor (NELF) are the key players for Pol II pausing, although other factors contribute to the stability of paused Pol II [20, 32]. The mechanisms behind promoter proximal Pol II pausing remain elusive, with new aspects being continually revealed.

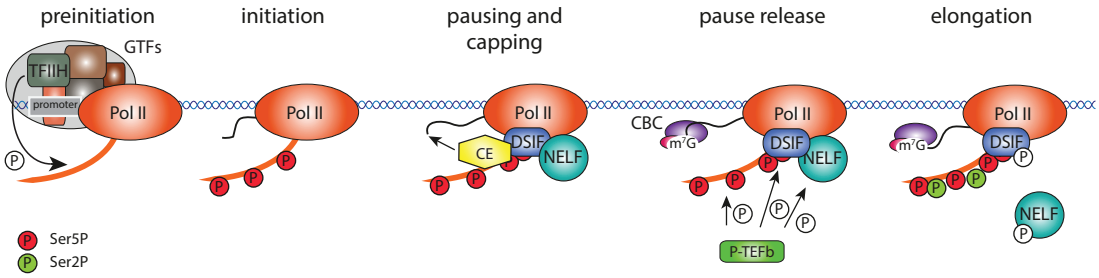


Figure 1.1 – Promoter proximal pausing of RNA Polymerase II

After forming of the preinitiation complex, the CTD of RNA polymerase II (Pol II) is phosphorylated at Ser5 by the transcription factor TFIIH. Ser5P leads to initiation of transcription through the recruitment of elongation factors. After 20-40 nts Pol II pausing is induced by binding of DSIF and NELF. The capping enzymes (CEs), which cap the nascent transcript at its 5′ end, are recruited by Ser5P and DSIF. Phosphorylation of NELF, DSIF and the CTD at Ser2 by P-TEFb induces the release of the paused Pol II and productive elongation. Phosphorylated NELF dissociates from the complex and phosphorylated DSIF acts as positive elongation factor. At this stage also other positive elongation factors are recruited. The m^7G cap structure is bound by the cap binding complex CBC, which accompanies the RNA through the following maturation and transport processes.

A critical factor essential and sufficient for the release of paused Pol II is P-TEFb [169]. P-TEFb consists of cyclin T and CDK9, a kinase that phosphorylates the NELF-E subunit of NELF [60] and the CTR repeats of DSIF at Thr4 [226]. Phosphorylated NELF dissociates from the stalled complex and phosphorylated DSIF acts as elongation-stimulating factor [126, 207]. Subsequently, P-TEFb also phosphorylates Ser2 of the Pol II CTD [166]. The recruitment of P-TEFb seems to be the rate-limiting

step in the release of paused Pol II and is regulated through several protein factors and cellular signals [21, 169, 233].

The function of promoter proximal Pol II pausing is unclear. Hypotheses include that Pol II pausing enables rapid and/or synchronous transcription of specific genes and that pausing serves as checkpoint for coupling elongation and RNA processing for which the 5′ cap structure is important. The latter is supported by the fact that 5′ capping occurs when the nascent RNA extends from 20-30 nts in length and that Ser5P is bound by the CE stimulating its activity [52, 65]. Also DSIF enhances the recruitment and activity of the CE (by CE binding to the CTR of Spt5) [7, 124, 131, 165, 214], while the binding of CE rescues transcriptional repression by NELF [131]. Further evidence for a coupling of promoter proximal pausing, capping and transcription elongation is provided by the discovery that the nuclear cap binding complex (CBC), which binds the cap structure of the nascent transcript (see Sec. 1.4), promotes the recruitment of P-TEFb, thus linking capping with transcription elongation [120].

1.3.3 *5′ end Capping*

The co-transcriptionally addition of a 7-methylguanosine (m^7G) cap structure to the 5′ end of the nascent transcript is a common feature all Pol II transcripts share. At a length of 20-30 nts a m^7G is connected via a 5′-5′ triphosphate bridge to the first nucleotide of the nascent transcript [150, 190]. The addition of m^7G is catalysed by three enzymatic activities (Figure 1.2A). First, the RNA 5′ triphosphatase removes the terminal γ-phosphate of the transcript, generating a diphosphate 5′ end. Then the RNA guanylyltransferase (GTase) catalyses the addition of guanosine monophosphate (GMP) to the 5′ diphosphate, followed by the addition of the methyl group to the N-7 amine of the guanosine cap by the guanine-7-methyltransferase to form the cap 0 structure. In higher eukaryotes, post-transcriptionally the first and second nucleotide can be further methylated at the 2´-*O* position of the ribose generating the cap 1 and cap 2 structures, respectively (Figure 1.2).

These unique caps of Pol II transcripts are recognised and bound by specific processing factors. In the nucleus the nuclear cap binding complex CBC (see Sec. 1.4) immediately binds the 5′ m^7G cap structure, protecting the RNA from degradation and accompanying the transcripts though various processing steps that are promoted by the cap.

The specificity of the m^7G cap to Pol II transcripts arises from the recruitment of the CE by the CTD of Pol II. The mammalian GTase directly binds to the Ser5 and Ser2 phosphorylated Pol II CTD [83]. Furthermore, Ser5P binding seems to stimulate the formation of the enzyme-GMP intermediate and serves as allosteric activator of the CE [65, 83].

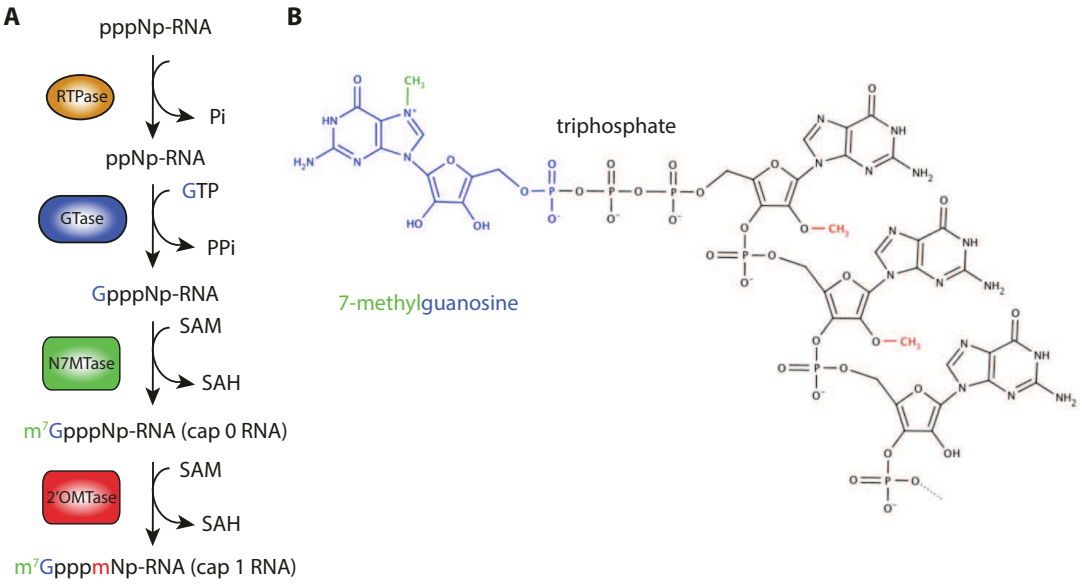


Figure 1.2 – The 5′ 7-methylguanosine cap

A: 5′ cap synthesis. The RNA 5′-triphosphatase (RTPase) hydrolyses the γ-phosphate from the 5′ end of the nascent transcript. Then the guanylyltransferase (GTase) transfers a GMP to the newly formed end and the guanine-7-methyltransferase (N7MTase) methylates the guanosine on position 7. 2′-O-methylation at the first nucleotide by a 2´-O-ribose methyltransferase (2′OMTase) leads to the generation of the cap 1 structure. SAM: S-adenosylmethionine, SAH: S-adenosylhomocysteine.
B: 5′ cap structure. The 7-methylguanosine (m^7G) is linked via a 5′-5′ triphosphate bridge to the RNA. Additional methylation of the first nucleotides results in the cap 1 and cap 2 structure. Scheme adapted from [167].

1.3.4 *Splicing*

Most precursor messenger RNA (pre-mRNA) transcripts and precursor of long intergenic/intervening non-coding RNAs (lincRNAs) contain sequences (introns) that are intervening the mature RNA sequences (exons). For proper RNA maturation the removal of the introns is essential. The respective process is called RNA splicing and comprises cleavage at conserved sequences and ligation of the exons by several steps. In addition, alternative splicing enables the arrangement of selected exons and gives rise to distinct RNAs from the same precursor RNA. RNA splicing is catalysed by the spliceosome, which is composed of small nuclear ribonucleoproteins (snRNPs) and many additional co-factors [96, 218]. For correct pre-mRNA splicing the cap structure as well as CBC are required. It was shown that the co-transcriptional assembly of the spliceosome relies on the interaction of CBC with U4/U6.U5 tri-snRNP [162]. Also a model including the interaction between CBC and splicing factors has been proposed [120].

1.3.5 *3′ end Formation*

1.3.5.1 *mRNA 3′ end Processing*

The 3′ end of a pre-mRNA is not defined by its translational stop codon as Pol II transcribes also the 3′ untranslated region (UTR) after the stop codon. For proper maturation the nascent pre-mRNA transcript must be cleaved and polyadenylated at its 3′ end. This 3′ end processing of pre-mRNAs plays an important role at many stages. It enhances the stability of mRNAs [217, 220], promotes mRNA export from the nucleus to the cytoplasm [199, 204] and translation of mRNAs into proteins [29, 185, 217, 220]. Furthermore, 3′ end processing is linked to transcription and splicing [82, 171, 173, 234] and the poly(A) tail is necessary for transcriptional termination [22, 171, 172].

The 3 ′end cleavage of pre-mRNAs is based on the recognition of *cis*-acting elements in the 3′ UTR of the pre-mRNA by a complex machinery. Following cleavage, pre-mRNAs, except histone replication-dependent transcripts (see Sec. 1.3.5.2), acquire a polyadenylated tail at the newly formed 3′ end. In mammals, the cleavage site is defined by the following *cis*-elements. 10-30 bases upstream of the cleavage site located is the hexamer AAUAAA polyadenylation signal (PAS), which is recognised by the five subunit cleavage and polyadenylation specificity factor (CPSF) complex [30]. About 30 bases downstream of the cleavage site lies a U- or GU-rich element called DSE (distal sequence element), that is recognised by the heterotrimeric cleavage stimulatory factor (CstF). Often additionally a U-rich element, which is recognised by the heterodimeric cleavage factor I_m (CFI_m) [19, 203] and enhances the efficiency of cleavage and polyadenylation, is present in one or more copies upstream of the cleavage site [89]. Another auxiliary element can be located downstream of the cleavage site and is generally G-rich [10, 30]. The cleavage site itself is not strictly conserved but cleavage preferentially occurs after a CA dinucleotide [132].

The mammalian 3′ end processing machinery contains more than 14 proteins comprising CPSF, CstF, CFI_m and $CFII_m$ subcomplexes as well as poly(A) polymerase (PAP), poly(A) binding protein (PABP), Symplekin and Pol II CTD. Some of these factors are also involved in transcriptional initiation, transcriptional termination or splicing. The cleavage is catalysed by the enzyme CPSF73, a subunit of CPSF, leading to a 3′ hydroxyl and 5′ phosphate end, respectively [45, 133, 181]. After cleavage PAP, which interacts with CPSF, adds 150-250 adenylate residues in a template independent manner to the 3′ end. However, in mammals PAP is not only required for polyadenylation, but also for cleavage. The new, short poly(A) tail is bound by PABP [208] enhancing the affinity of PAP for the RNA [100]. After around 250 nts the interaction between PAP and CPSF weakens, which halts polyadenylation and thus determines the length of the poly(A) tail [109, 209].

The mammalian Pol II CTD is necessary for cleavage [142] and its phosphorylation stimulates *in vitro* 3′ end processing [182]. The CTD interacts with CPSF via CPSF160 [41, 142] and CstF via CstF50 [142]. However, already during transcription CPSF associates via CPSF30 with the body of Pol II and signals the polymerase to terminate transcription after passing the TTATTT PAS sequence [153].

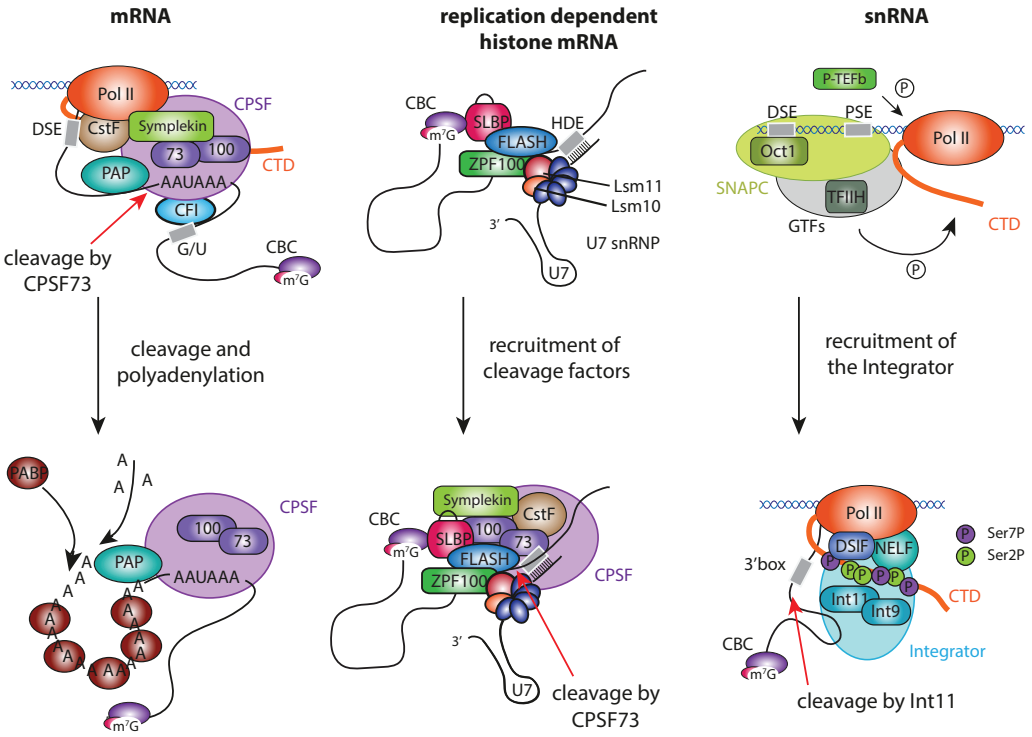


Figure 1.3 – 3'end processing of mRNA, replication dependent histone mRNA and snRNA

Models displaying the main factors as well as structural and sequence elements of the RNA involved in 3' end processing of mRNA, replication dependent histone mRNA and snRNA. For more details see Sec. 1.3.5.

1.3.5.2 *Histone mRNA 3´ end Processing*

During S phase, the genomic DNA gets replicated and histone proteins must be produced in large amounts. As the different histone types must be synthesised in a fixed ratio, their synthesis must be coordinated with the replication of DNA. Therefore, their transcription and expression are strictly regulated in a cell cycle dependent manner and differ from the processing of other pre-mRNAs. Replication dependent histone pre-mRNAs lack introns and apart from 5' capping (see Sec. 1.3.3), the 3' end processing is their only maturation process, which includes a cleavage but no polyadenylation.

In contrast to other pre-mRNAs the cleavage site of replication dependent histone pre-mRNAs is defined by structural and sequence elements [44]. A conserved 16 nts stem-loop lies in the 3' UTR and a purine-rich sequence called histone downstream element (HDE) is located 12-19 nts downstream of the cleavage site. The highly conserved stem-loop, consisting of a six nts stem and four nts U-rich loop, lies 24-70 nts downstream of the stop codon and five nts upstream of the cleavage site. The stem-loop is bound by the stem-loop binding protein (SLBP) [134, 212] and the HDE interacts with U7 snRNP by base pairing with the U7 snRNA (Figure 1.3). Then the cleavage factor is recruited. It comprises proteins of the CPSF (CPSF73, CPSF100 and FIP1) and Symplekin, which are also part of the complex that processes 3' end

polyadenylated pre-mRNAs. CPSF73 seems to be the endonuclease cleaving both histone pre-mRNAs [45] and polyadenylated pre-mRNAs [133, 181]. Additionally, CBC [156], NELF [156], ZFP100 [43], FLICE-associated huge protein (FLASH) [194, 229], arsenite-resistance protein 2 (ARS2) [71, 183] and U2 snRNP [59] are involved in replication dependent histone pre-mRNA 3′ end processing. However, the exact mechanism leading to proper histone mRNA processing instead of polyadenlyation and the function of the individual proteins remains poorly understood.

1.3.5.3 *snRNA 3´ end Processing*

The processing of precursor small nuclear RNA (pre-snRNA) is comparable to replication dependent histone pre-mRNA processing (see Sec. 1.3.5.2). Following 3′ end endonucleolytic cleavage in the nucleus, the RNAs are not polyadenylated and contain a terminal 3′ stem loop structure [67]. However, the endonucleolytic cleavage is mediated by a machinery containing the multidomain Integrator complex and the 3′ end of pre-snRNAs undergoes additional trimming in the cytoplasm.

For U snRNA transcription an U snRNA specific promoter containing a distal sequence element (DSE) and a proximal sequence element (PSE) are required [23]. These sequences recruit snRNA specific transcription factors (Oct1 and snRNA activation protein complexes (SNAPCs)) in addition to the general transcription factors (GTFs) (TFIIA, TFIIB, TFIIE, TFIIF, TFIIH) [148]. Furthermore, snRNA genes contain a conserved sequence called 3′ box 9-19 nts downstream of the mature end of the U snRNA [79, 80, 230].

In the current model of pre-snRNA co-transcriptional 3′ end processing (Figure 1.3) the cyclin-dependent kinase (CDK)7 of TFIIH phosphorylates the Pol II CTD on Ser5 and Ser7 [2, 68] upon transcription initiation of U snRNA genes. Ser7P recruits the putative Ser5 phosphatase RNA polymerase II associated protein 2 (RPAP2), which leads to the association of first Integrator subunits. During elongation Ser5P is dephosphorylated by RPAP2 and CDK9, the kinase of P-TEFb, phosphorylates Ser2, allowing the remaining Integrator subunits to bind. After the 3′ end box is transcribed Integrator cleaves the nascent pre-snRNA, which might be recognised via its stem-loop. Integrator is a conserved multisubunit complex of which two domains, Int9 and Int11, show homology with CPSF100 and CPSF73, respectively. This homology leads to the assumption that Int11 is the endonuclease of the Integrator complex [3].

Recent studies showed that Integrator is recruited not only to loci of snRNA genes but also binds to the 3´ end of replication dependent histone pre-mRNAs and promoter proximal sites at protein coding transcripts that Ser7P is not required for its recruitment [193]. Furthermore, Integrator seems to be involved in Pol II transcription initiation, pause release, elongation as well as termination at mRNA encoding genes [63, 193, 196]. At mRNAs, snRNAs and replication dependent histone mRNAs, Integrator co-localises with NELF and DSIF and its recruitment is DSIF-dependent [193]. Thus, DSIF and NELF as well as Integrator appear to be involved in the proper RNA maturation of different Pol II RNA classes [228]. However, to understand their function within the distinct RNA processing, further investigations are required.

1.3.6 *Nuclear RNA Export*

Regulation of gene expression also occurs at the level of RNA transport. For example the separation of transcription and translation allows the proper processing and quality control of new mRNAs. The mature RNAs are exported from the nucleus to the cytoplasm through the nuclear pore complex (NPC). This requires the assembly of transport factors to the transcripts to form export-competent RNPs. Part of these RNPs are the mature RNA, CBC, export adaptor and export receptor proteins. The adaptor proteins associate with RNA already during RNA transcription and processing and serve as RNA mark to recruit the respective export receptor. Export receptors mediate the RNPs transport through the NPC. There are two major heterodimeric export receptors Nxf1-Nxt1 and CRM1-Ran leading to distinct export pathways. Most mRNAs are exported in Nxf1-Nxt1 dependent manner, whereas rRNAs, snRNAs are exported via the CRM1-Ran pathway (Figure 1.4), which also exports nuclear export signal (NES) containing proteins [57, 61, 159, 195].

1.3.6.1 *mRNA Export*

In metazoa the transcription-export (TREX) complex is recruited to the nascent mRNPs by the splicing machinery, whereas in yeast the transcription machinery recruits the TREX complex [33, 135, 198]. In humans the TREX complex contains ALY, UAP56, Tex1 and the multisubunit THO complex. The THO complex and UAP56, an RNA helicase essential for splicing, assemble first on mRNA. UAP56 then recruits the adaptor protein ALY to mRNA, followed by UAP56 dissociation from the complex in an ATP-dependent manner [47, 200]. This leads to the recruitment of Nxf1-Nxt1 [47, 77, 197–199]. Then the mRNP is rearranged and ALY dissociates from the transcript, allowing Nxf1 to bind the mRNA [77]. Subsequently, Nxf1-Nxt1 shuttles the transcript through the NPC [98, 187, 189]. Notably, also other adaptor proteins have been reported like UAP56 interacting factor (UIF) and Chtop [27, 78]. Also CBC was described to link transport receptors and mRNA [33, 74, 157] and CBP80, the bigger subunit of the CBC heterodimer, directly interacts with ALY [33]. Furthermore, some mRNAs seem to be exported via the CRM1-Ran pathway [38, 103] (Sec. 1.3.6.2).

1.3.6.2 *U snRNA Export in Metazoans*

Typically, snRNAs are shorter than mRNAs ranging from 200-300 nts [136] and associate specifically with the snRNA export adaptor PHAX (phosphorylated adaptor for RNA export) [140, 158]. In contrast, longer transcripts are bound by the heterogeneous nuclear ribonucleoprotein (hnRNP) C1/C2, which replaces PHAX from mRNAs but not from U snRNAs [140]. Within the pre-export complex containing CBC, PHAX and the transcript, PHAX binds to both CBC and snRNA. Subsequently, PHAX is phosphorylated by the casein kinase 2 (CK2), leading to the recruitment of the export receptor CRM1 together with the small GTPase Ran [158]. Interestingly, phosphorylated PHAX alone does not interact with chromosomal region maintenance 1 (CRM1) [158]. In the cytoplasm GTP hydrolysis by Ran and dephos-

phorylation of PHAX by protein phosphatase 2A (PP2A) leads to the disassembly of the U snRNA export complex [105, 158].

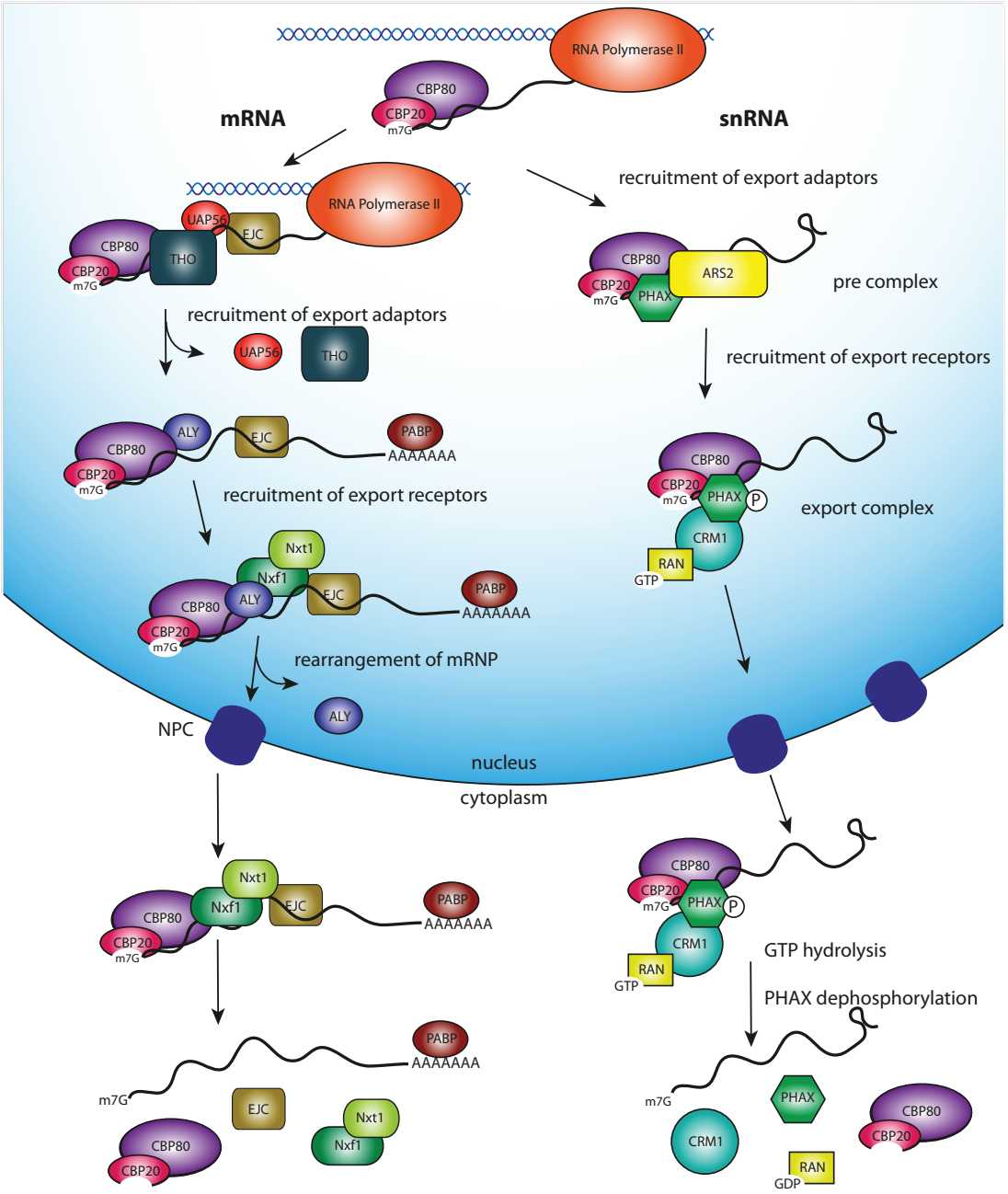


Figure 1.4 – Nuclear RNA export

Model illustrating the main factors involved in the nuclear export of mRNAs and snRNAs. mRNAs are mainly exported via the Nxf1-Nxt1 pathway (see Sec. 1.3.6.1), while snRNAs require the export adaptor PHAX and the export receptor CRM1-Ran (see Sec. 1.3.6.2).

1.3.7 *Nuclear Quality Control*

Pol II transcripts undergo various maturation processes including 5′ capping, splicing, 3′ end processing and packing into export-competent RNPs. Although these processes are very efficient, errors can occur within each step leading to the production of aberrant RNAs and/or the formation of aberrant RNPs. As these irregular RNAs/RNPs can interfere with the normal RNA metabolism and protein production, the cell needs to destruct the non-functional transcripts. Additionally, also the accumulation of unwanted RNAs and transcriptional by-products can disrupt the RNA metabolism and must be prevented. To ensure that only functional RNAs are further processed and exported from the nucleus, different nuclear quality control pathways are linked to the individual processing steps. The nuclear exosome plays a central role in the degradation of dysfunctional RNA molecules and thus in the regulation of the RNA metabolism [4, 34, 146, 147]. The eukaryotic exosome is a conserved ribonucleolytic complex consisting of 10 or 11 subunits with 3′-5′ exo- and endo-nucleolytic activities [147]. To achieve its full activity and to target specific and correct substrates co-factors need to assemble to the RNAs [116, 154]. The main adapter complex required for the degradation of RNAs is the TRAMP (Trf4/Air2/MTR4 polyadenylation) complex, containing the MTR4 RNA helicase that unwinds the transcript [110, 202, 224]. However, for promoter upstream transcripts (PROMPTs), enhancer RNAs (eRNAs) and pre-snRNAs, the nuclear exosome targeting (NEXT) complex consisting of RBM7, ZCCHC8 and MTR4 links the RNAs via MTR4 to the exosome [128]. Additionally, the poly(A) tail exosome targeting (PAXT) complex, containing ZFC3H1, PABPN1 and MTR4 promotes the exosomal degradation of longer and more mature RNAs than the NEXT complex [144]. Interestingly, all complexes require MTR4 and the interaction of ZCCHC8 and ZFC3H1 with MTR4 is mutually exclusive, leading to the selection of the NEXT or PAXT pathway [144]. Moreover, both complexes, NEXT and PAXT, interact with the CBC-ARS2 complex via ZC3H18 [6, 144]. Also additional pathways including DICER and XRN2 are under discussion to target RNAs for degradation [39] and further studies are needed to understand how the different co-factors are recruited to target the distinct dysfunctional transcript classes for degradation.

1.4 THE NUCLEAR CAP BINDING COMPLEX - CBC

Immediately after 5' capping the highly conserved cap binding complex CBC, consisting of CBP80 (NCBP1 - 90 kDa) and CBP20 (NCBP2 - 20 kDa), specifically binds to the m^7G cap of the nascent Pol II transcript [94]. Depending on the capped RNA, different nuclear RNA processing and transport machineries are recruited and CBC seems to interact directly or indirectly with some of these components (e.g. ALY [33], CTIF [102], hnRNP C [140], NELF [156]). Therefore, CBC is involved and plays an important role in many RNA processing events including pre-mRNA splicing [94], RNA 3´ end processing [56, 156], nonsense-mediated decay [86, 118], mRNA and snRNA nuclear exports [33, 93] and miRNA processing [72, 113, 184]. Additionally, it was shown that CBC is implicated in the pioneer round of translation [92]. However, the role of CBC in helping the distinct RNAs to select their appropriate nuclear RNA maturation pathways is poorly understood.

The crystal structure of CBC in complex with m^7GpppG (PDB: 1H2T [139]) shows that CBP20 with its RNA recognition motif (RRM) binds the cap structure (Figure 1.5A, B), although the interaction of both CBP80 and CBP20 is necessary for cap binding [93, 94, 99]. Furthermore, CBP20, especially the residues involved in cap binding, is highly conserved (Figure 1.5C). Cap binding leads to stabilisation and proper folding of around 50 residues of CBP20 surrounding the RRM domain [139]. Hence, apo-CBC [137] and CBC bound to the cap [138, 139] can be distinguished by their overall conformation. CBP80 contains three linked domains, each of them containing consecutive helical hairpins that resemble the MIF4G (middle domain of eIF4G) domain (Figure 1.5A).

In addition to the nanomolar affinity and high specificity of CBC for the m^7G cap structure [222], CBP20 as well as CBP80 seem to directly interact with RNA [12, 24, 161]. However, the significance of these interactions remains elusive.

In the cell CBC shows a predominantly nuclear localisation, although it is exported from the nucleus to the cytoplasm together with the capped RNA through the nuclear pore (see Sec. 1.3.6). A bipartite nuclear localisation signal (NLS), important for the correct localisation of CBC, lies in the N-terminus of CBP80 [95, 205] and CBP20 is presumed to be co-imported with CBP80 via the importin-α/importin-β pathway [122]. Additionally, the release of the cap from CBC in the cytoplasm is promoted by binding of CBC to the nuclear import machinery (importin-α and importin-β) [40, 70].

Although CBC has been shown to be important for the proper processing of many RNAs, the depletion of CBC in *S. cerevisiae* or *A. thaliana* leads to a growth deficient phenotype, but is not lethal [1, 58, 201]. Knock down of CBP80 in mammalian cells results in a reduced cell proliferation [156] and depletion of CBP20 impacts the export of mRNAs leading to accumulation of poly(A) RNAs in the nucleus [33, 64].

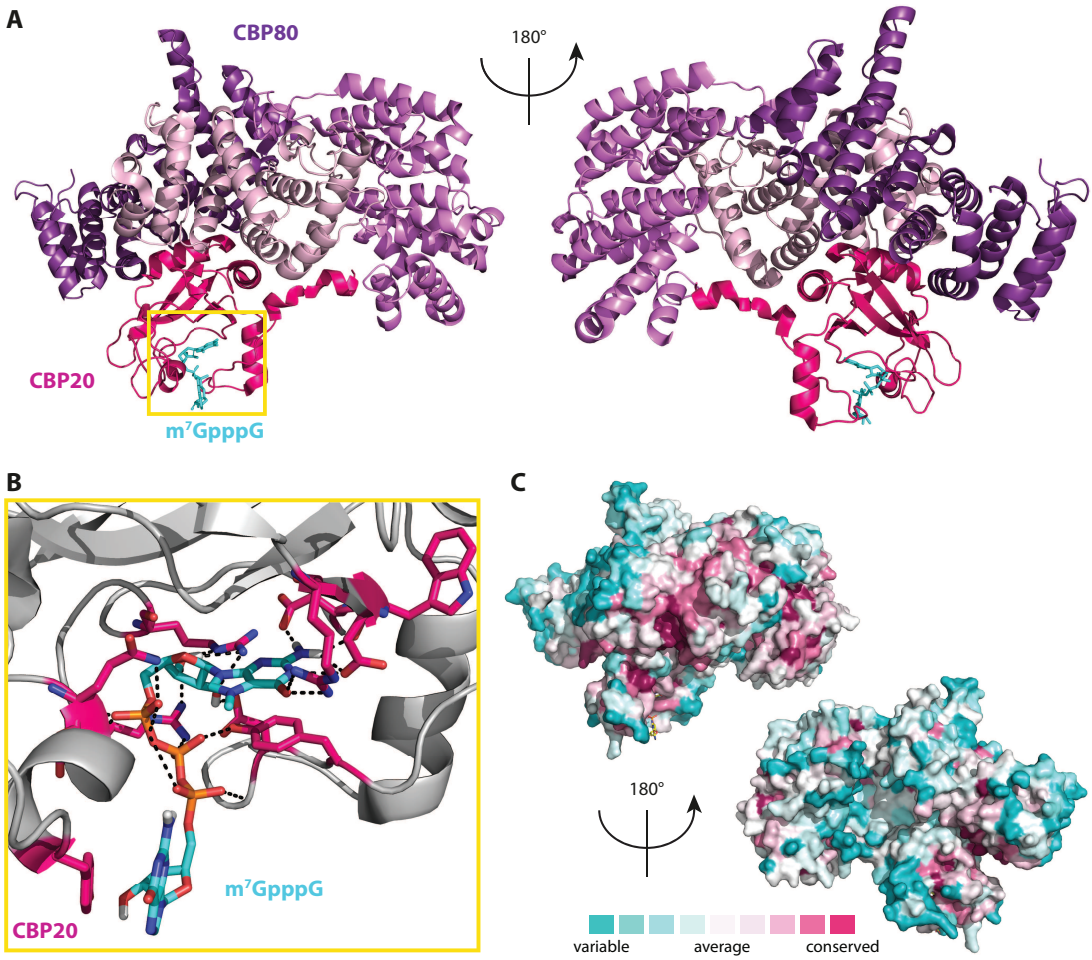


Figure 1.5 – Structural overview of the nuclear cap binding complex CBC

A: Crystal structure of CBC bound to m^{7}GpppG (PDB: 1H2T [138]). CBP80 is made up of three helical domains (purple shades) and CBP20 (pink) is binding m^{7}GpppG (cyan).
B: Close up view of the cap binding site within CBP20.
C: Surface conservation of human CBC displayed on the m^{7}GpppG-CBC crystal structure (PDB: 1H2T [138]).

1.5 THE NUCLEAR CAP BINDING PROTEIN 3 - NCBP3

Recently, another putative cap binding protein the nuclear cap binding protein 3 (NCBP3), also known as ELG protein, has been described [64]. Initially NCBP3 was identified as component of messenger ribonucleoprotein (mRNP) particles involved in splicing [145]. NCBP3 is a protein of 73 kDa with a predicted canonical RNA recognition motif (RRM) in its N-terminal region (residues 126-187). A N-terminal construct of NCBP3 containing the RRM domain (NCBP3$^{1\text{-}282}$) bound m^{7}GTP in the micromolar range (K_D~15 µM) [64]. Additionally, mutations in the conserved WLDD motif of the RRM domain, which is supposed to be located in the RNA binding

grove, resulted in loss of this interaction. It has been postulated that NCBP3 directly interacts with CBP80 to form an alternative cap binding complex [64]. In addition, proteins of the TREX complex, the exon junction complex (EJC) as well as mRNAs seem to associate with NCBP3. Therefore, NCBP3 is possibly involved in the export of mRNAs [64]. However, its precise function and the underlying mechanism remain unclear.

1.6 THE PHOSPHORYLATED ADAPTOR FOR RNA EXPORT - PHAX

PHAX [158] is a conserved, metazoan specific protein of 44 kDa. It contains two NLS, a NES and an RNA binding domain (RBD) [188]. Two major phosphorylation sites were identified to be important for the formation of the nuclear export complex comprising CBC, PHAX, CRM1 and RanGTP [105]. Although phosphorylation of PHAX is required for interaction with CRM1 and RanGTP, this trimeric complex alone is unstable and requires further stabilization by U snRNA and CBC [158]. PHAX is phosphorylated in the nucleus by CK2 and dephosphorylated in the cytoplasm by PP2A [105]. This compartmentalised phosphorylation/dephosphorylation system is important for the U snRNA export (see Sec. 1.3.6.2). Recently, it was also shown that ARS2 binds to the CBC-PHAX complex and to snRNAs stimulating their 3' end processing [73]. Furthermore, PHAX is involved in the intranuclear transport of snoRNAs together with CBC and CRM1 [15, 16, 111]. Thus, depletion of PHAX leads to incorrect nuclear localisation of U snRNAs and snoRNAs and hinders the nuclear export of U snRNAs [119, 213].

Until now, structural information is only available for the RBD of PHAX (Figure 1.6). Nuclear magnetic resonance (NMR) studies showed that the human PHAX RBD only acquires a ternary structure upon binding to single stranded (ss) RNA [151]. Furthermore, PHAX shows a weak affinity to ssDNA (150 µM) and binds ss RNA sequence unspecific in the micromolar range (4.2-6.5 µM) [151]. The structure of PHAX RBD bound to AUCG reveals an entirely α-helical fold (PDB: 2XC7 [151])

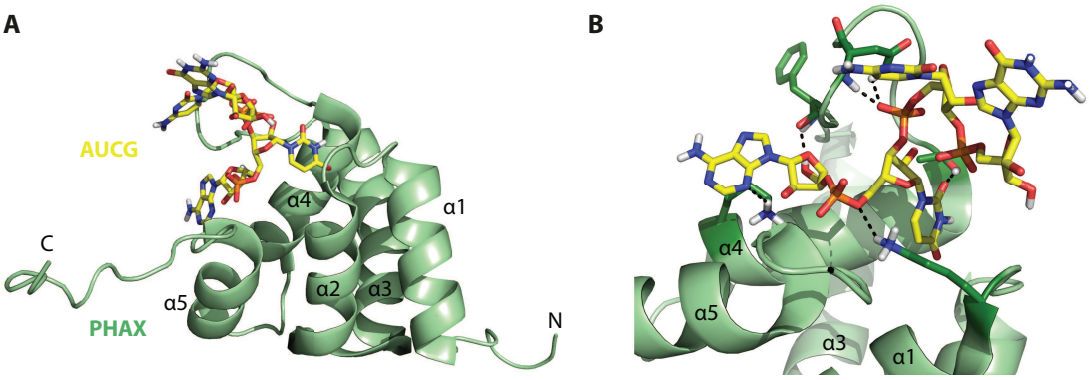


Figure 1.6 – Solution structure of PHAX RNA binding domain bound to ssRNA

A: NMR structure of PHAX RBD in complex with the ssRNA AUCG (PDB: 2XC7 [151]).
B: Close up view of the RNA binding site showing the interaction of PHAX with the RNA sugar-phosphate backbone.

(Figure 1.6A). The interaction of the RNA and PHAX is mainly mediated by the sugar-phosphate backbone, explaining the non-sequence-specific binding (Figure 1.6B).

1.7 THE ARSENITE-RESISTANCE PROTEIN 2 - ARS2

ARS2, also known as serrate RNA effector molecule homolog SRRT, is an 107 kDa conserved metazoan protein [5, 69, 170, 219] initially identified as protein involved in arsenite resistance [180]. Later studies demonstrated that ARS2 is essential for cell proliferation and development in *Drosophila*, zebrafish and mouse [5, 69, 72, 219]. ARS2 shows a predominantly nuclear localisation [219] and its NLS is supposed to be part of the N-terminal 147 amino acids [114, 160]. ARS2 was shown to be part of several distinct RNA complexes involved in miRNA biogenesis [72, 160], snRNA export [73], exosomal degradation [6, 144] and RNA 3′ end processing [71, 73, 104]. ARS2 binds directly to CBC as well as to CBC-PHAX [73]. Furthermore, ARS2 also seems to interact with proteins involved in different RNA processing pathways such as FLASH [104], ZC3H18 [6] and DROSHA [72], thereby linking CBC bound transcripts to the distinct processing machineries. Knockdown of ARS2 in mammalian cells leads to defects in 3′ end processing of certain transcripts such as histone mRNAs and snRNAs and in reduced degradation of short-lived transcripts like PROMPTs [6, 71, 73]. The precise biological function of ARS2 is largely unknown, especially on molecular level, since only the crystal structure of the *A. thaliana* ARS2 homolog SERRATE is available (PDB: 3AX1 [129]), which adopts an unusual protein structure (Figure 1.7). Remarkably, the metazoan ARS2 contains an additional highly conserved RRM domain, which is absent in plants [160]. Five isoforms of human ARS2 are described.

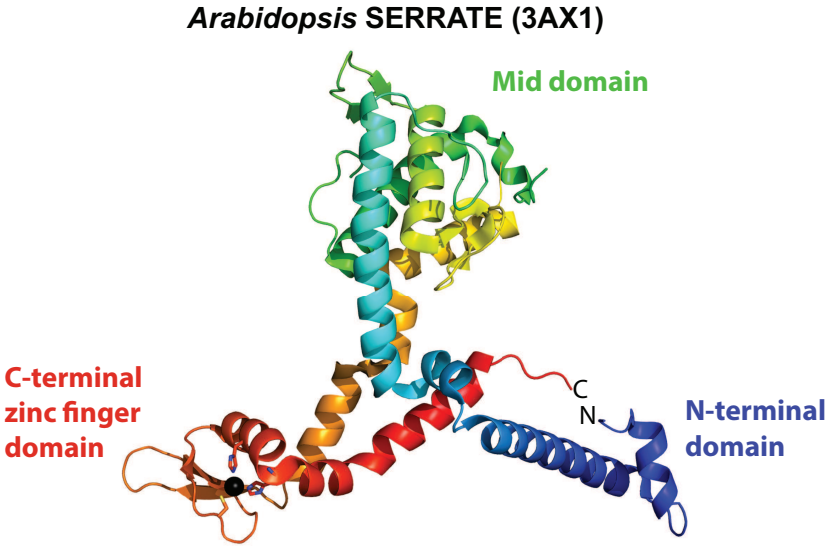


Figure 1.7 – Crystal structure of ARS2 plant homolog SERRATE

The *A. thaliana* SERRATE crystal structure (PDB:3AX1 [129]) is rainbow coloured with the N-terminus in blue and the C-terminus in red.

They are produced by alternative splicing and show either a deletion of a single residue (isoform 3 - E388), the replacement of five residues by a serine (isoform 2 - 776-780 to S), the combination of both (isoform 4) or the deletion of the single residue 387 combined with the deletion of a residue 196-231 (isoform 5). However, the existence of isoform 5 as protein has not been confirmed experimentally.

Recent studies analysed ARS2 by mutagenesis to map possible RNA and protein-protein interaction sites, suppose that ARS2 is either associated with DROSHA and primary microRNA (pri-miRNA) or FLASH and replication dependent histone mRNA [160]. Furthermore, it has been reported that the interaction between FLASH and ARS2 is required for S phase progression [104], demonstrating a direct link between the two proteins and their relevance within the cell cycle. The ARS2-FLASH interaction is mediated by only 13 out of 1982 amino acids of FLASH, called FLASH-ARS2-binding (FARB) region [104]. Until now the actual binding site of FLASH and other proteins as well as RNA on ARS2 has not been identified.

1.8 THE NEGATIVE ELONGATION FACTOR - NELF

Another protein complex shown to be involved in 3′ end processing of replication dependent histone mRNA is NELF, which only exists in metazoa [155]. NELF consists of the four subunits NELF-A, NELF-B, NELF-C (or splice variant NELF-D) and NELF-E [155, 227]. The NELF complex was originally identified to mediate Pol II pausing together with DSIF [177, 227] (see Sec. 1.3.2). Recently, it has been shown that NELF-A and NELF-C form the core of the NELF complex. NELF-A contains a Pol II binding site, whereas all other NELF subunits associate with RNA [206]. The RRM domain of NELF-E binds to a specific RNA motif, called NELF-E binding element (NBE), which is enriched in promoter proximal regions in *Drosophila* [174]. Thus, the binding of NELF-E to NBE is thought to regulate the promoter proximal pausing of Pol II in a transcript specific manner [163].

Interestingly, NELF-E directly binds to CBC and it was shown that NELF-E's C-terminal residues (NELF-E$^{360\text{-}380}$) are important for this interaction [156]. Depletion of either CBP80 or NELF-E lead to the accumulation of aberrant poly-adenylated histone mRNAs in HeLa cells, indicating a common role in correct 3′ end processing of replication-dependent histone mRNA [156]. Within the last years the role of NELF in proper 3´ end processing has been extended to snRNAs. It was shown that NELF and DSIF (via respectively NELF-A and Spt5) interact with the multiprotein Integrator complex, which is responsible for 3´ end processing of pre-snRNAs [196, 228] (see Sec. 1.3.5.3).

1.9 AIM OF THE WORK

The nuclear cap binding complex CBC plays a major role in the maturation of many different classes of Pol II transcripts. It directly binds the nascent transcripts and serves as binding surface for processing factors. These interactions seem to be crucial for the correct processing and export of Pol II transcripts. However, the molecular mechanisms that give rise to the individual complexes remain unknown. Thus, the goal of this doctoral thesis was to elucidate the role of CBC in sorting different families of Pol II transcripts to the appropriate nuclear processing pathway. Therefore, different CBC complexes including CBC-PHAX, CBC-ARS2 and CBC-NELF-E were investigated. For a better understanding of the arrangement of the complexes dif-

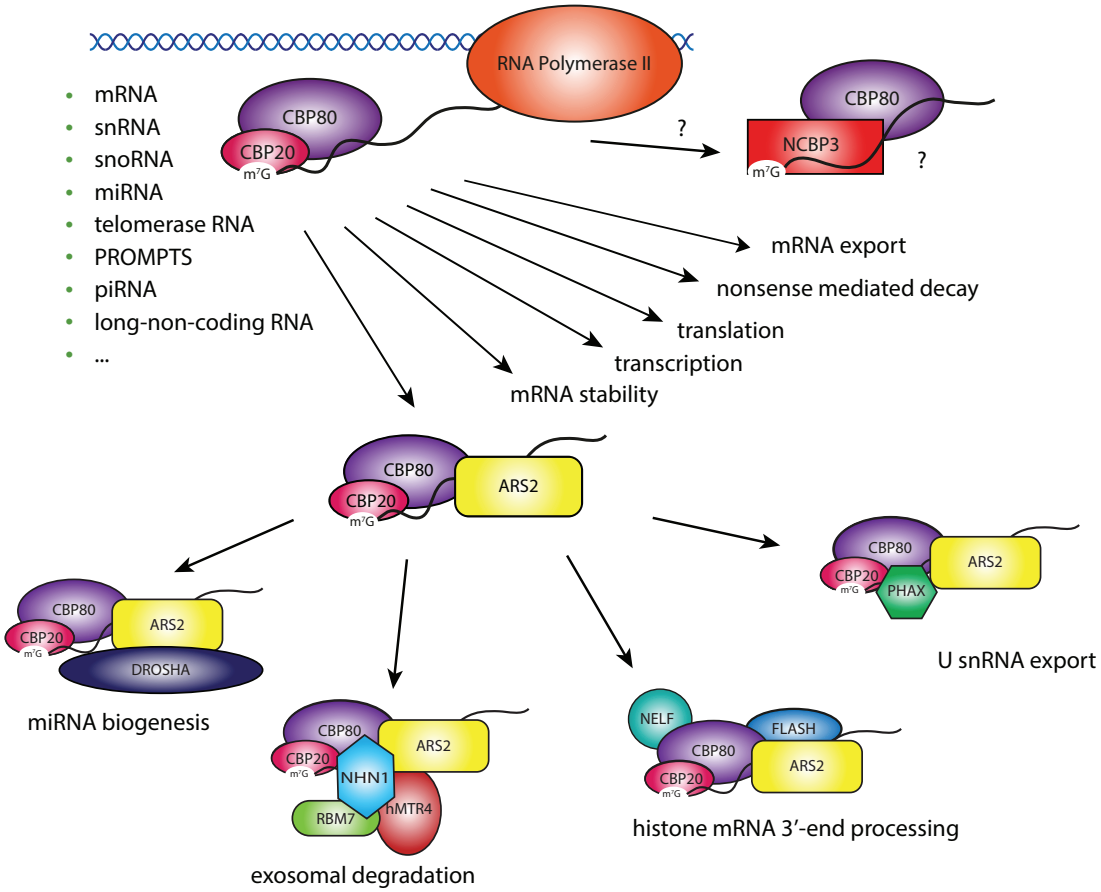


Figure 1.8 – Model of RNA Polymerase II transcript pathways with focus on CBC-ARS2

All Pol II transcripts are bound by CBC via their m7G cap before entering different processing and transport pathways. Depending on the RNA ARS2 or even ARS2 plus additional proteins are recruited to the complex. However, the molecular basis of the recruitment and formation of the individual processing complexes remains elusive. Additionally, another putative cap binding complex consisting of NCBP3 and CBP80 with unknown function has been described recently [64].

ferent biophysical, structural and biochemical methods were used to analyse the protein-protein interaction.

Apart from CBC, also ARS2 associates with different RNAs (e.g. miRNAs, histone mRNAs, snRNsA and mRNAs) and interacts with proteins involved in distinct RNA processing pathways (e.g. FLASH [104], DROSHA [72] and ZC3H18 [6]). To get insight into the role of ARS2 and its interactions atomic resolution structural information is required. Currently, there is only a crystal structure of the homologous protein SERRATE from *A. thaliana* available, which exhibits a unique 'walking man' (body, leading and lagging leg) topology [129]. However, metazoan ARS2 differs from SERRATE in having a number of insertions including a putative RRM domain of unknown function ([160]). Thus, another goal was to obtain structural information of ARS2 and to investigate its interaction with RNA and different RNA processing machineries (e.g. PAXT, NEXT, DROSHA and FLASH).

Additionally, the recent description of NCBP3 and CBP80 as alternative nuclear cap binding complex was analysed to understand and to learn more about its function within RNA processing.

2 MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 *Antibodies*

- DIS Alexa Fluor 488 anti-BrdU antibody (BioLegend) - Mouse IgG1, κ
- Alexa Fluor 647 anti-GFP antibody (BioLegend) - Rat IgG2a, κ

2.1.2 *Bacteria*

- *E. coli* Rosetta 2 (DE3) (Novagen): F-ompT hsdSB(rB- mB-) gal dcm (DE3) pRARE2 (Cam^R)
- *E. coli* DH10 Bac (Invitrogen): F-mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ-rpsL nupG/pMON14272/pMON7124
- *E. coli* TOP10 (Invitrogen): F-mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 araD139 Δ(ara, leu)7697 galU galK rpsL (Str^R) endA1 nupG

2.1.3 *Chemicals*

Chemicals were purchased from the following companies: Euromedex, Fluka, Sigma-Aldrich, CARLO ERBA Reagents, Merck, MP Biomedicals, Roche, Roth.

2.1.4 *Cell Lines*

- *Homo sapiens* embryonic kidney (HEK 293T/17) cells (ATCC)
- *Spodoptera frugiperda* (Sf21) insect cells (Gibco)
- *Trichoplusia ni* (High Five) insect cells (Gibco)

2.1.5 *Composition of Solutions*

All buffers and solutions were prepared with ultrapure water (Millipore systems) and filtered through a 0.22 μm membrane filter. Buffers for chromatography at an ÄKTA System were additionally degassed before use. The composition of all media and buffers used in this work are listed in Table A.1 and Table A.2 in Appendix A.1.

2.1.6 *Cross-linker*

- disuccinimidyl suberate DSS -H12/D12 (Creative Molecules Inc.)

2.1.7 *Enzymes*

Enzymes used in this thesis are listed in Table 2.1.

Table 2.1 – Enzymes

ENZYME	MANUFACTURER
Calf Intestinal Phosphatase (CIP)	NEB
DNA Ligation Kit, Mighty Mix	Takara
Phusion High Fidelity DNA Polymerase	NEB
Polymerase GoTaq Green Master Mix	Promega
PrimeSTAR Max DNA Polymerase	Takara
Restriction Endonucleases	NEB
RNase A, DNase and protease free	Thermo Fisher Scientific
Sequencing Grade Modified Trypsin	Promega
T7 Polymerase	EMBL Heidelberg, Germany
His-tagged TEV protease	EMBL Heidelberg, Germany
Lysyl Endopeptidase (Lys-C), Mass Spectrometry Grade	Wako

2.1.8 *Instruments*

Instruments used in this study are listed in Table A.3 in Appendix A.2.

2.1.9 *Kits*

Kits used in this study are listed in Appendix A.3 Table A.4.

2.1.10 *Oligonucleotides*

DNA-oligonucleotides were ordered from Thermo Fisher Scientific (Waltham, United States). RNA-oligonucleotides and FAM-labeled RNA-oligonucleotides were obtained from iba (Göttingen, Germany).

2.1.11 *Peptides*

Peptides and FAM-labeled peptides were bought from GenScript (Piscataway, United States).

2.1.12 *Plasmids*

General plasmids used as basis within this study are listed in Table 2.2. Table A.5 in Appendix A.4 contains all plasmids constructed for protein expression used within this work.

Table 2.2 – Plasmids

PLASMID	RESISTANCE	MANUFACTURER	DESCRIPTION
pcDNA3.1	Amp	Invitrogen	one multiple cloning site, for mammalian cell line expression
pETM11	Kan	EMBL	N-terminal His tag followed by TEV cleavage site, for bacterial expression
pETM30	Amp	EMBL	N-terminal His tag followed by GST tag and TEV cleavage site, for bacterial expression
pETM41	Amp	EMBL	N-terminal His tag followed by MBP tag and TEV cleavage site, for bacterial expression
pETM11 SUMO	Amp	EMBL	N-terminal His tag followed by SUMO tag and TEV cleavage site, for bacterial expression
pETDuet	Amp	Novagen	two multiple cloning sites (one with N-terminal His tag followed by TEV cleavage site), for bacterial expression
pET-15b	Amp	Takara	two multiple cloning sites (one with N-terminal His tag followed by TEV cleavage site), for bacterial expression
pFastBac	Amp	Life Technologies	one multiple cloning site, for baculovirus expression
pFastBac HTb	Amp	Life Technologies	N-terminal His tag followed by TEV cleavage site, for baculovirus expression
pFastBac Dual	Amp	Life Technologies	two multiple cloning sites (one with N-terminal His tag followed by TEV cleavage site), for baculovirus expression

2.1.13 *Saccharomyces cerevisiae Strains*

The yeast strains used within this study are listed in Table 2.3.

Table 2.3 – Yeast strain genotypes

STRAIN	GENOTYPE	REFERENCE
CG1945	MATa, ura3-52, his3-200, ade2-101, lys2-801, trp1-901, leu2-3, 112, gal4-542, gal80-538, cyhr2, LYS2::$GAL1_{UAS}$-$GAL1_{TATA}$-HIS3, URA3::GAL4 17-mers (x3)-$CyC1_{TATA}$-lacZ	[55]
Y187	MATα, ura3-52, his3-200, ade2-101, trp1-901, leu2-3, 112, gal4Δ, met–, gal80Δ, URA3::$GAL1_{UAS}$-$GAL1_{TATA}$-lacZ	[76]

2.1.14 *Transfection Reagents*

- X-tremeGENE HP DNA Transfection Reagent (Roche Life Science)
- polyethylenimine (PEI), branched M_W 25 kDa (Sigma-Aldrich)

2.2 MOLECULAR BIOLOGY METHODS

2.2.1 *Antibiotics*

After transformation of a resistance encoding plasmid the following antibiotics were used at the specified concentrations:

- ampicillin (100 mg/ml)
- chloramphenicol (30 mg/ml)
- gentamycin (7 mg/ml)
- kanamycin (50 mg/ml)
- streptomycin (50 mg/ml)
- tetracycline (10 mg/ml)

2.2.2 *Polymerase Chain Reaction*

DNA fragments for cloning were amplified with Phusion High-Fidelity DNA Polymerase or PrimeStar Max (Takara) according to the standard procedures [186]. For colony polymerase chain reaction (PCR) GoTaq polymerase (Promega) was used.

2.2.3 *Agarose Gel Electrophoresis*

DNA fragments were separated depending on their size in 1-2% (w/v) agarose EDTA gels supplemented with 1:20 000 SYBR Safe DNA Gel Stain (Invitrogen). DNA was visualised using a UV lamp.

2.2.4 *DNA Restriction*

All restriction digestions were performed with Type II endonucleases (NEB) according manufacturer's protocols.

2.2.5 *DNA Purification*

After PCR and restriction digestions DNA fragments were purified using either the QIAquick PCR Purification Kit (Qiagen), the QIAquick Gel Extraction Kit (Qiagen) or the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) according to manufacturer's protocol.

2.2.6 *DNA Ligation*

Purified insert and plasmid were ligated using the DNA Ligation Kit, Mighty Mix (Takara). Plasmid and insert were mixed in an 1:3 molar ratio and incubated for 1 h at 16°C.

2.2.7 *In-Fusion Cloning*

To insert one or multiple DNA fragments at once into a plasmid the In-Fusion HD Cloning Kit (Clontech) was used. Plasmid and insert containing a 15-20 bp overlap at their ends were mixed in an 1:1.1 molar ratio and incubated with the In-Fusion HD Enzyme Premix for 15 min at 55°C prior transformation into *E. coli* (see Sec. 2.2.12).

2.2.8 *Plasmid Preparation*

Plasmids were isolated and purified using the QIAprep Spin Miniprep Kit (Qiagen) or NucleoSpin Plasmid Kit (Macherey-Nagel) according to manufacturer's instructions.

2.2.9 *DNA and RNA Concentration Determination*

To estimate the concentrations of single stranded (ss) DNA and ssRNA oligonucleotides the absorbance at 260 nm was measured with a NanoDrop spectrophotometer (Thermo Fisher Scientific).

2.2.10 *Site-directed Mutagenesis*

Mutations were introduced using site directed mutagenesis [84, 176]. Two overlapping primers containing the mutation of interest were used for PCR (see Sec. 2.2.2). After DpnI digestion the DNA was transformed into *E. coli* (see Sec. 2.2.12).

2.2.11 *Preparation of Chemically Competent E. coli*

Chemically competent *E. coli* were prepared using a standard rubidium chloride method. LB media supplemented with the respective antibiotics was inoculated with 1% (v/v) of an overnight culture and grown at 37°C until the $OD_{600\,nm}$ reached 0.6. Bacteria were pelleted (6500 g, 10 min, 4°C), resuspended gently in solution CC1 and incubated for 1 h on ice. After another round of centrifugation (6500 g, 10 min, 4°C) the bacteria were resuspended in buffer CC2 and incubated 15 min on ice. Then the bacteria were aliquoted and flash frozen in liquid nitrogen before stored at -80°C.

2.2.12 *Transformation of Chemically Competent E. coli*

DNA was transformed into *E. coli* via heat shock. Briefly, chemically competent bacteria were thawed on ice and incubated with the DNA of interest. After 10-20 min they were incubated for 1 min at 42°C following 1-5 min incubation on ice. 1 ml of LB medium was added and the bacteria were incubated for 30-60 min at 37°C before plating on the appropriate selective media and incubation at 37°C overnight.

2.2.13 *Transformation of Saccharomyces cerevisiae*

Appropriate yeast colonies were resuspended in 500 µl 0.1 M lithium acetate and centrifuged (2500 g, 2 min, RT). After resuspending the pellet in 100 µl 0.1 M lithium acetate, 10 µl 10 mg/ml herring sperm DNA, 1-2 µg plasmid DNA and 700 µl 0.1 M lithium acetate, 40% (w/v) PEG in Tris-EDTA Buffer were added, mixed and incubated 30 min at 30°C. Then 88 µl DMSO were added and the cells thermo shocked for 7 min at 42°C. The cells were pelleted (2500 g, 2 min, RT) and washed once with water before plated on the corresponding plates.

2.2.14 *Yeast Two-Hybrid Assay*

For the yeast bridged two-hybrid assay pACT2 and p422 (ADE2 multicopy) plasmids were transformed into Y187 and pAS2 plasmids were introduced into CG1945 haploid *Saccharomyces cerevisiae*. Strains were crossed for 4 h at 30°C and diploids were plated on triple- and quadruple-selective media (-Ade –Leu -Trp or -Ade -Leu -Trp -His). After 3 days at 30°C cell growths was assessed visually.

2.2.15 *Bacmid Production*

1-2 µg of plasmid DNA encoding the gene of interest were transformed into *E. coli* DH10 Bac by heat shock as described in Sec. 2.2.12. To enable DNA recombination the culture was incubated 4-6 h at 37°C. Afterwards different dilution were streaked out on kanamycin/tetracycline/gentamycin /IPTG/Bluo-Gal plates. To exclude any contamination, the next day white colonies were picked and restreaked. For the bacmid preparation two 2 ml overnight cultures of two individual white clones were pelleted

(2200 g, 5 min, RT). For the following steps buffers from the QIAprep Spin Miniprep Kit (Qiagen) were used. The pellet was resuspended in 300 µl Buffer P1 and transferred to an 1.5 ml tube before adding 300 µl of Buffer P2. The tube was gently inverted and incubated for 2 min at room temperature. Afterwards 300 µl of Buffer N3 were added and the tube was inverted ten times. The suspension was centrifuged (18000 g, 10 min, RT) and the supernatant was transferred to a new tube followed by centrifugation (18000 g, 5 min, RT). In a new tube 700 µl of 100% isopropanol were added to the supernatant and gently mixed. The DNA was pelleted (18000 g, 10 min, RT) and washed twice with 70% (v/v) ethanol (18000 g, 5 min, RT).

2.2.16 *Protein Constructs*

An overview of all in this thesis used constructs encoding for the protein of interest can be found in Table A.5 in Appendix A.4.

2.3 CELL BIOLOGY METHODS

2.3.1 *Insect Cell Culture*

Sf21 and High Five insect cells were cultured shaking at 27°C at a density of $0.5 \cdot 10^6$ cells/ml in their individual medium.

2.3.2 *Baculovirus Production*

After isolation of the bacmid (see Sec. 2.2.15) the ethanol was removed and the bacmid was first resuspended in 20 µl water before further dilution with 200 µl SF21 medium. 10 µl of X-tremeGENE transfection reagent (Roche Life Science) were added to 100 µl of SF21 medium and then added together to the DNA containing tube. After incubation for 10 min 0.5-$1 \cdot 10^6$ Sf21 cells/well were transfected with 150 µl of the mixture in a 6-well plate. After 72 h the supernatant (V_0) was removed and stored at 4°C. To obtain the V_1 25 ml of Sf21 at $0.5 \cdot 10^6$ cells/ml were infected with 3 ml of V_0. Cells were diluted with medium to maintain the density below $1 \cdot 10^6$ cells/ml. 72 h after proliferation arrest the V_1 was harvested by pelleting the cells (10000 g, 10 min, RT) and stored at 4°C.

2.3.3 *Mammalian Cell Culture*

HEK 293T/17 cells were grown in high glucose Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% (v/v) fetal bovine serum (FBS). Cells were maintained at 37°C and 5% (v/v) CO_2 and split every 2-3 days. To split the cells the medium was removed and cells were washed twice with PBS before incubation with 0.01% (w/v) trypsin-EDTA for 3 min. Then cells were resuspended and diluted in whole medium before seeded in new flasks.

2.3.4 *Transfection of Mammalian Cell Lines*

Mammalian cells were transfected using PEI in a DNA:PEI ratio of 1:2 (Table 2.4). One day before transfection cells were split to the desired density.

Table 2.4 – Transfected DNA mixtures

DISH	DNA	PEI
6-well plate	1 µg	2 µl
T25 flask	2.5 µg	5 µl
T75 flask	7 µg	14 µl

2.3.5 *Cell Cycle Analysis*

For bromodeoxyuridine (BrdU) cell cycle experiments, $0.5 \cdot 10^5$ HEK 293T/17 cells were seeded on T25 flasks and the next day transfected with the desired plasmids. 24 h post transfection, cells were transferred onto a T75 flask to allow growth. The next day, cells were given a 30 min BrdU pulse and fixed with 60% (v/v) ethanol overnight. Cells were incubated with 2 M hydrochloric acid for 20 min at 37°C. After centrifugation (4 min, 400 g, RT) cells were resuspended in 0.1 M sodium borate. After washing the cells twice with PBS (4 min, 400 g, RT), cells were counted and labeled with mouse Alexa Fluor 488 conjugated anti-BrdU antibody (BioLegend) and rat Alexa Fluor 647 conjugated anti-GFP antibody (BioLegend) in Labeling Buffer for 1 h at room temperature. Cells were washed three times (4 min, 400 g, RT) and resuspended in Labeling Buffer supplemented with 3 µM 4',6-diamidino-2-phenylindole (DAPI). A total of 10^4 GFP-positive events per sample were collected on a BD LSRFortessa flow cytometer at the Flow Cytometry Core Facility EMBL Heidelberg.

2.3.6 *Immunofluorescence*

For immunofluorescence $3 \cdot 10^5$ cells/well were seeded onto a glass coverslip in a 6-well plate. The next day cells were transfected with 1 µg of DNA. 48 h after transfection cells were washed twice with PBS and incubated with 4% (w/v) PFA in PBS for 15 min at room temperature in the dark followed by a PBS wash and 15 min incubation with 0.1% (v/v) Triton X-100 in PBS. Then the coverslips were washed two more times with PBS, mounted using Fluoroshield Mounting Medium with DAPI (abcam) onto microscope slides and sealed using nail polish. Slides were kept in the dark at 4°C until analysed by confocal microscopy (Leica TCS SP5).

2.4 EXPRESSION AND PURIFICATION OF PROTEINS

2.4.1 *Bacterial Protein Expression*

For bacterial protein expression the constructs of interest (Appendix A.4 Table A.5) were transformed into *E. coli* Rossetta 2 and generally expressed overnight at 18°C after induction with 0.4 mM IPTG at an $OD_{600\ nm}$ of 0.6-0.9. CBP20, GST and PHAX full length were expressed at 30°C for 6 h. The cells were harvested by centrifugation (800 g, 10 min, 10°C), flash frozen in liquid nitrogen and stored at -20°C before protein purification.

2.4.2 *Protein Expression in Insect Cells*

For protein expression High Five insect cells were cultivated in Express Five Serum-free Medium (Invitrogen) supplemented with 18 mM L-glutamine and 100 µg/ml penicillin and streptomycin. Cells were transfected at $0.5 \cdot 10^{6}$ cells/ml with 0.1-0.5% (v/v) virus V_1 to stop cell division after 24 h. After additional 48 h the cells were harvested by centrifugation (800 g, 10 min, 10°C), flash frozen in liquid nitrogen and stored at -20°C before protein purification.

2.4.3 *Seleno-methionine Labeling in Bacteria*

M9 medium supplemented with 1 mM magnesium sulfate, 0.1 mM calcium chloride, 20% (w/v) glucose, 2 mg/ml thiamine and appropriate antibiotics was inoculated with 0.1% (v/v) LB preculture and grown at 37°C overnight. 50 ml of this preculture were used to inoculate each final liter of supplemented M9 medium. At an $OD_{600\ nm}$ of 0.7-0.9 the temperature was lowered to 18°C. The dissolved amino acids (100 mg Lys, 100 mg Phe, 100 mg Thr, 50 mg Ile, 50 mg Leu, 50 mg Val and 50 mg seleno-Met per liter) were added and after additional 30 min the protein expression was induced using 0.3 mM IPTG. The bacteria were harvested by centrifugation (6500 g, 15 min, 10°C) and stored at -20°C.

2.4.4 *Cell Lysis*

Bacteria and insect cells were lysed using a sonicator. The pellets were resuspended in Lysis Buffer and sonicated on ice for 3.5 min (10 sec on/25 sec off) at an intensity of 75%. Afterwards cell debris and aggregations were separated by centrifugation (60000 g, 1 h, 10°C).

2.4.5 *Ni Affinity Chromatography*

After lysis and centrifugation, the cleared supernatant was loaded on Ni-Sepharose (GE Healthcare) equilibrated with appropriate buffer in a BioRad gravity flow column. For the purification of CBC, first cleared His-tagged CBP20 lysate was loaded

on the Ni-Sepharose. After washing with Lysis Buffer and the cleared CBP80 supernatant was applied to the column. After washing with at least 10 column volumes of Lysis Buffer supplemented with 20 mM imidazole, the His-tagged protein was eluted with 300 mM imidazole in Lysis Buffer. Wash and elution fractions were analysed by SDS-PAGE (see Sec. 2.5.2) and protein containing fractions pooled and further purified.

2.4.6 *Tag Removal*

To cleave the affinity tags from the construct of interest all fusion proteins carried the specific recognition site for TEV protease. After elution from the affinity chromatography fractions were pooled and overnight dialysed against 50-400 mM sodium chloride, 10 mM β-mercaptoethanol, 5% (v/v) glycerol, 20 mM HEPES pH 7.8 in the presence of 1 mg TEV protease (EMBL Heidelberg) per 50 mg recombinant protein. Afterwards the protein solution was applied again to the Ni affinity chromatography and the flow through kept for further purification.

2.4.7 *Ion Exchange Chromatography*

Before loading the sample on the column the salt concentration was adjusted by dialysis or diluting the sample to allow the sample to bind to the resin. After washing away all unbound components the protein was eluted using a sodium chloride gradient on an ÄKTA Protein Purification System (ÄKTA Prime, Basic, or Purifier, GE Healthcare).

2.4.8 *Size Exclusion Chromatography*

Size exclusion chromatography (SEC) was used in the final step of purification, for buffer exchange, for checking the protein quality, for interaction studies and for reconstitution of protein-complexes.

Depending on the size of the protein or protein-complex a Superdex 200 PC3.2/30, Superdex 75 10/300, Superdex 200 10/300 or a Superose 6 10/300 was used on an ÄKTA System. Up to 1 ml of sample was injected on the column equilibrated in the buffer of choice. Flow rates were typically 0.4 ml/min for 10/300 columns on an ÄKTA Protein Purification System and 0.05 ml/min for PC3.2/30 columns on an ÄKTA Micro system.

2.4.9 *Concentration of Protein Solutions*

Protein solutions were concentrated using Amicon Ultra Centrifugal Filters (Merck Millipore) with a cut off below half of the molecular weight of the protein. In case of protein precipitation during concentration the sodium chloride concentration was increased up to 500 mM.

2.4.10 *Protein Storage*

For long-term storage or shipment purified protein was divided into small aliquots and flash-frozen in liquid nitrogen. Aliquots were stored at -80 °C and shipped on dry ice.

2.5 BIOCHEMICAL AND BIOPHYSICAL METHODS

2.5.1 *Protein Concentration Determination*

Protein concentrations were determined using the calorimetric Bio-Rad Protein Assay, an assay based on the absorbance shift of Coomassie brillant blue G-250 [17].

2.5.2 *Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis*

The presence and purity of proteins were analysed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). Depending on the size of the proteins between 6-18% (w/v) SDS-polyacrylamide gels were cast. Samples were mixed with SDS Sample Buffer and electrophoresis was conducted with 1x SDS Running Buffer (200-300 V for 30-60 min). To estimate the size of protein bands a molecular weight standard was used. Gels were stained with Coomassie (InstantBlue, Expedeon) for 5-10 min followed by washing and destaining with water. To detect low concentrations of protein on a SDS-PAGE gel as from a single crystal the Silver Stain Plus Kit (Bio-Rad) was used following manufacturer's instructions.

2.5.3 *Identification of Proteins from SDS-PAGE Gels by Mass Spectrometry*

After electrophoresis (see Sec. 2.5.2) SDS-PAGE gels were fixed with 45% (v/v) methanol and 10% (v/v) acetic acid for 15-60 min. After rinsing the gel twice for 5 min with water the gel was stained with Coomassie blue solution for 1-3 hours. The gel was washed twice with water before destained overnight with 45% (v/v) methanol and 10% (v/v) acetic acid. After washing with water, the gel was stored in 1% (v/v) acetic acid. The bands were excised from the gel, trypsin digested and identified by mass spectrometry (EMBL Proteomics Core Facility Heidelberg).

2.5.4 *Limited Proteolysis*

To identify stable protein subcomplexes and protein-protein complexes the individual proteins or complexes were treated with different proteases. Therefore, the proteins were incubated with trypsin (1:1000 (w/w)), chymotrypsin (1:500 (w/w)), elastase (1:500 (w/w)) or papain (1:1000 (w/w)) in 50 mM HEPES pH 7.7, 150 mM sodium chloride for up to 2.5 h at room temperature or over night at 4°C before analysis by SDS-PAGE or gel filtration followed by SDS-PAGE (see Sec. 2.5.2).

2.5.5 *Isothermal Titration Calorimetry*

To determine thermodynamic parameters of protein-protein interactions the isothermal titration calorimetry (ITC) was used. In the ITC experiment one protein solution in the syringe is titrated into the cell containing the other protein. In case of molecular interaction heat is either released or absorbed. Measurements of this difference allow the determination of the standard Gibbs free energy change ΔG, the equilibrium binding constant K_a and the dissociation constant K_D, the enthalpy change ΔH, the entropy change ΔS and the stoichiometry n of the interaction. Correlation between the parameters is given by: $\Delta G = -RTlnK_a = RTlnK_D = \Delta H - T\Delta S$ (with R as universal gas constant and T as temperature in degrees Kelvin).

ITC measurements were performed at 25°C with a MicroCal200 calorimeter (Malvern). Prior to measurements, proteins were dialysed overnight in the same buffer to exclude buffer differences. The titration was performed by stepwise addition of the 5-20 times higher concentrated sample in the syringe to the protein in the sample cell at a stirring speed of 800 rpm. As control the solution in the syringe was titrated into buffer to subtract possible effects due to dilution during titration. Data analysis was carried out with the MicroCal ITC Origin software suite (Malvern). The integrated heat values were fitted to the "single binding site" model and iterated until the chi squared value reached the minimum. K_D values and standard deviations listed within this study represent the average from at least two independent experiments.

2.5.6 *Fluorescence Polarisation Assay*

Fluorescence polarisation binding assays were performed with N-terminal FAM-labeled ARS2$^{845\text{-}871}$ (ARS2$^{845\text{-}871}$(FAM)) and NELF-E$^{354\text{-}380}$ (NELF-E$^{354\text{-}380}$(FAM)) peptides. 40 nM peptides were titrated with increasing concentrations of proteins in Buffer A. For the competition assay CBC-ARS2 and CBC-NELF-E complexes were preformed in the presence of 25 µM m^{7}GTP using 40 nM ARS2$^{827\text{-}871}$(FAM) peptide and 60 nM CBC and 40 nM NELF-E$^{354\text{-}380}$(FAM) peptide with 1 µM CBC. These preformed complexes were titrated with NELF-E$^{244\text{-}380}$, ARS2$^{827\text{-}871}$ and PHAX$^{103\text{-}327}$, respectively. For the displacement assay between PHAX and NELF-E, PHAX$^{103\text{-}327}$ was titrated to a premixture of 40 nM ARS2$^{827\text{-}871}$(FAM) peptide, 60 nM CBC and 270 nM NELF-E$^{244\text{-}380}$ in the presence of 25 µM m^{7}GTP.

For RNA-binding studies FAM-labeled RNA was used at a concentration of 37.5 nM and titrated with increasing amounts of protein in RNA Binding Buffer.

Fluorescence polarisation was measured at 25 °C with a microplate reader (BMG LABTECH) using an excitation wavelength of 495 nm and an emission wavelength of 515 nm. For analysis the polarisation value for the peptide alone was subtracted from the measurements and the K_D values derived by curve fitting using GraphPad Prism version 5.00 and 7.00 for Windows (GraphPad Software, La Jolla, California, United States).

2.5.7 *Protein Cross-Linking Combined with Mass Spectrometry*

Purified protein complexes were cross-linked at 0.5 mg/ml with 2 mM DSS (Creative Molecules Inc.) at 35°C at 600 rpm for 30 min. To quench the reaction 0.1 volume quenching solution (250 mM ammonium bicarbonate, 10 M urea) were added and incubated additional 10 min at 35°C and 600 rpm. Then additional 0.8 volume quenching solutions were added and the sample twice sonicated for 5 s. After addition of DTT to a final concentration of 10 mM and incubation for 30 min at 30°C and 600 rpm, iodoacetamide was added to a final concentration of 15 mM and incubated 30 min at room temperature in the dark. Proteins were digested by addition of Lys-C (Wako) (1:100, 4 h at 37 °C) and following dilution of urea to 1.5 M with trypsin (Promega) (1:50, over night at 37°C). 2% (v/v) trifluoroacetic acid (TFA) was added and incubated for 30 min at 37°C. After centrifugation (10 min, 12 000 g, RT) samples were loaded on Macro SpinColumns (Harvard Apparatus) (4 min, 1000 g, RT). After four washes with 500 µl 5% (v/v) acetonitril (ACN), 0.1% (v/v) formic acid (FA) (2 min, 1000 g, RT) the peptides were eluted three times with 250 µl in 50% (v/v) ACN, 0.1% (v/v) FA (4 min, 1000 g, RT) and lyophilised in a Speed-Vac. Samples were reconstituted in 20 µl water/ACN/TFA (70:30:0.1) and subjected to SEC (Superdex Peptide PC3.2/30). Fractions of interest were lyophilised and sent to the Proteomics Core Facility (EMBL Heidelberg) for mass spectrometric analysis. For the identification of peptides the xQuest/xProphet software was used [117, 211]. Cross-linked peptides were filtered with a delta score <0.95 and a false discovery rate <0.05.

2.5.8 *m^7GTP-Sepharose Pull Down*

To pull down specifically cap-binding proteins m^7GTP-Sepharose (GE Healthcare) affinity chromatography was used. The purified proteins were incubated with prequilibrated resin rotating at 10°C for at least 30 min. For the competition assay or to test interaction with other proteins the resin bound cap binding complex was incubated with additional protein before washing with buffer and eluting using 2x SDS Sample Buffer. The presence of proteins of interest in the flow though, wash and elution was analysed by SDS-PAGE (see Sec. 2.5.2).

2.5.9 *GST Pull Down*

Purified GST-tagged proteins or GST-tagged proteins from precleared lysate were immobilised using Glutathione Sepharose (GE Healthcare) affinity resin before incubation with possible interacting partners in Buffer A. After extensive washing with Buffer A (400 g, 2 min, 10°C) the complexes were eluted using 10 mM reduced Glutathione or 2x SDS Sample Buffer. Presence of proteins of interest in the elution was analysed by SDS-PAGE (see Sec. 2.5.2).

2.5.10 *EGFP Affinity Purification Followed by Mass Spectrometry*

Frozen cell pellets were thawed on ice, resuspended in HEK 293 Lysis Buffer supplemented with protease inhibitor tablets and 1 µg/ml RNAse H and 30 min incubated on ice. After centrifugation (16000 g, 20 min, 4°C) the supernatant was precleared for 1 h at 4°C with 100 µl preequilibrated Sepharose A beads (GE Healthcare) before incubation with 25 µl GFP-Trap (ChromoTek) beads at 4°C for 2 h. After washing the beads three times with 500 µl HEK 293 Lysis Buffer (400 g, 2 min, 10°C) the protein was eluted using 50 µl 0.1 M glycine pH 2.3 followed by a second elution with 30 µl. The eluted protein was neutralised with 0.1 M Tris-HCl pH 8.5 and flash frozen in liquid nitrogen. Within the Proteomics Core Facility (EMBL Heidelberg) the samples were digested and labeled with tandem mass tags (TMTs) before analysed by mass spectrometry. Also further analysis of the obtained data was performed by the Proteomics Core Facility.

2.6 PROTEIN CRYSTALLIZATION AND DATA COLLECTION

2.6.1 *Protein Crystallization*

Initial screening experiments were performed by the high-throughput crystallisation (HTX) platform at EMBL Grenoble using the six pre-made screens (Classic Suite, JCSG+, PACT premier, PEGs-I, Salt Grid, Wizard I&II). Conditions leading to crystals were manually reproduced and further optimised in terms of precipitant, salt and protein concentrations as well as screened for possible additives. Crystals were flash-frozen in well solution supplemented with 20% (v/v) glycerol prior X-ray diffraction analysis.

The CBC-ARS2$^{827\text{-}871}$ complex was prepared by mixing CBC with an excess of ARS2$^{827\text{-}871}$ in the presence of 1 mM m^{7}GTP and subjecting the mixture to gel filtration (120 mM sodium chloride, 5 mM β-mercaptoethanol, 20 mM HEPES pH 7.8). The complex was concentrated to 8 mg/ml using Amicon ultra centrifugal filters (Millipore). The best crystals were obtained in mother liquor containing 0.1 M sodium acetate pH 5, 8% (v/v) MPD and 0.1 M guanidine hydrochloride at 20°C. For the NELF-E complex m^{7}GTP or m^{7}GpppG together with NELF-E$^{360\text{-}380}$ or CBC-NELF-E$^{354\text{-}380}$ was added to CBC before setting up the drops. The best crystals of m^{7}GTP-CBC-NELF-E$^{360\text{-}380}$ were obtained by mixing 6 mg/ml CBC, 500 µM m^{7}GTP and 500 µM peptide in 120 mM sodium chloride, 1 mM TCEP, 20 mM HEPES pH 7.8. Crystals grew at 20°C in mother liquor containing 0.2 M lithium citrate and 20% (w/v) PEG 3350.

For ARS2 the best diffracting crystals were obtained at 4°C in mother liquor containing 0.2 M potassium citrate tribasic monohydrate, 20% (w/v) PEG 3350 (for ARS2 $^{147\text{-}270 + 408\text{-}763}$ and ARS2 $^{147\text{-}270 + 408\text{-}763\Delta 567\text{-}599\ GSGSGS}$); 1.2 M sodium dihydrogen phosphate, 0.8 M dipotassium hydrogen phosphate, 0.1 M CAPS pH 10.5, 0.2 M lithium sulfate (for ARS2 $^{171\text{-}270 + 408\text{-}763}$); 0.1 M Bis-Tris propane pH 6.5, 0.2 M lithium sulfate, 22% (w/v) PEG 3350 (for ARS2 $^{147\text{-}270 + 408\text{-}763\Delta 539\text{-}554\ GSA}$) and 0.2 M lithium sulfate, 20% (w/v) PEG 3550 (for ARS2$^{147\text{-}270+408\text{-}763}$ + FLASH$^{921\text{-}946}$).

2.6.2 *Data Collection*

X-ray diffraction data of single crystals were collected under cryo-conditions at the European Synchrotron Radiation Facility (ESRF) beamlines (Grenoble, France), namely, ID30A-1 (MASSIF-1), ID23-1 and ID29. On ID30A-1 datasets were recorded at a fixed energy of 12.835 keV (0.966 Å) with a Pilatus3 2M detector (Dectris). The tunable beamline ID23-1 was used with an energy of 14.200 keV (0.873Å) and is equipped with a PILATUS 6M-F detector (Dectris). On the tunable beamline ID29 datasets of seleno-methionine crystals were collected at 12.6578 keV (0.9795 Å) and all other datasets were collected at 12.600 keV (0.984 Å) with a PILATUS 6M-F detector (Dectris).

At ID23-1 and ID29 initial indexing and determination of the optimal data collection strategy were automatically calculated from four initial diffraction images using EDNA [90], MOSFLM [121], RADDOSE [164] and BEST [168] which are implemented in MXCube user interface [62].

2.7 BIOINFORMATICS AND COMPUTATIONAL METHODS

2.7.1 *Sequence Analysis and Alignments*

Multiple protein sequence alignments were generated with Clustal Omega [123, 143, 192] and visualised using ESPript [179]. Additionally, the CLC Sequence Viewer 6.9 (CLC Genomics Workbench 9.5.3) was used to align amino acid and DNA sequences.

2.7.2 *Secondary Structure and Protein Disorder Prediction*

Secondary structure and protein disorder prediction were performed with PrDOS [91] and JPred [46].

2.7.3 *Data Analysis*

ITC data were analysed with the MicroCal ITC-ORIGIN Analysis Software suite (Malvern) (see Sec. 2.5.5). Graphical presentations and fluorescence anisotropy curve fitting were performed with GraphPad Prism version 5.00 and 7.00 for Windows (GraphPad Software, La Jolla, California, United States). For the analysis and visualisation of the obtained flow cytometry data FlowJo V10 was used (TreeStar Software, Ashland, United States). After gating for single cells, cells with high EGFP staining were analysed for their BrdU incorporation in respect to their DNA amount. For the DNA/cell cycle analysis the univariate model was used to quantify the cells in the distinct cell phases. To compare the mutants with the wild type of all three experiments the One-way ANOVA followed by Dunnett's multiple comparisons test was performed using GraphPad Prism.

2.7.4 *Structure Analysis and Presentation*

Structure figures were drawn and root-mean-square deviation (RMSD) values were calculated with PyMOL (PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC). Conservation studies were performed using the ConSurf Server [8, 9, 25, 112]. The PISA server [108] was used to calculate macromolecular interfaces, LIGPLOT [210] to visualise ligand-protein interactions and BobScript [50, 51] to draw omit difference electron density maps.

2.7.5 *Structure Determination*

2.7.5.1 *Molecular Replacement*

X-ray diffraction data were indexed, integrated and scaled with XDS [97] and analysed using the CCP4i suite [221]. For CBC complexes the phase problem was solved by molecular replacement with PHASER [141] and using the known structure of CBC bound to m^7GpppG (PDB:1H2T [139]). After obtaining the ARS2 structure additional ARS2 data sets were solved by using the obtained structure as model for molecular replacement.

2.7.5.2 *Atomic Model Building, Refinement and Validation*

Refinement was performed with REFMAC5 [152] and BUSTER [18] with local non-crystallographic symmetry restraints. COOT [49] was used for model building and structure analysis. For structure validation MolProbity [31] and the PDB validation tools [175] were used.

2.7.5.3 *Phasing and Initial Model Building for ARS2*

For phasing of the ARS2$^{171\text{-}270\ +408\text{-}763}$ data a subset of four datasets was identified from a larger pool that maximised the anomalous signal while retaining acceptable merging R-values. The selection was performed with the help of Max Nanao using a genetic algorithm [231]. These datasets were merged in XSCALE and submitted to the CRANK2 pipeline in combination with the native dataset for substructure determination and phasing by single isomorphous replacement (SIR). The resulting map was used for automatic model building followed by iterative rounds of manual building and refinement (Sec. 2.7.5.2). The well-defined seleno-methionine positions were used as sequence markers.

3

RESULTS

3.1 BIOCHEMICAL CHARACTERISATION OF CBC COMPLEXES

To gain insight into the CBC-ARS2, CBC-PHAX, CBC-PHAX-ARS2 and CBC-NELF-E complexes, the different proteins were expressed and purified independently. Based on protein disorder predictions and described domains (Figure 3.1) different protein constructs were cloned in expression plasmids and the respective proteins were expressed in insect cells or bacteria. An overview of the constructs, plasmids and expression systems used in this study as well as the purification can be found in Section 2.1.12, Section 2.4 and Appendix A.4 Table A.5.

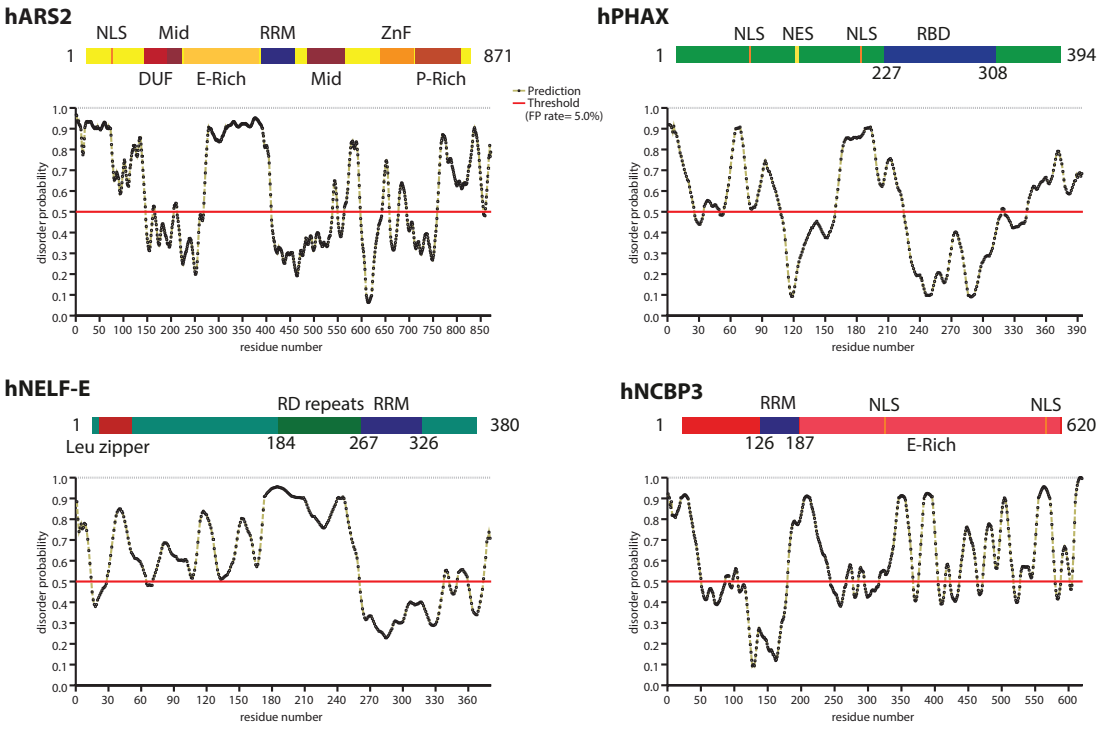


Figure 3.1 – Protein domain organisation and protein disorder prediction

Depiction of human (h) ARS2, PHAX, NELF-E and NCBP3 protein illustrating their reported domains and their disorder prediction of each protein. Region below threshold (red line) are considered to be ordered. FP: false positive rate of prediction.

3.1.1 *Protein Expression and Purification*

For bacterial protein expression pETM11 plasmids containing a T7 promoter, which originates from bacteriophage T7, showed a higher expression of the desired proteins than the pProEx HTb plasmid containing the synthetic trc promoter (data not shown). CBP20 showed only a weak bacterial expression from the pRSET A plasmid (data not shown). A higher yield was reached when the protein was expressed as GST-tagged protein from pETM30 (data not shown).

After plasmid optimisation CBP20, NCBP3, PHAX and NELF-E could be expressed as full length proteins, whereas soluble protein could only be obtained for a truncated construct of ARS2 (residues 147-871 or ARS2$^{147\text{-}871}$) lacking the predicted unstructured N-terminal region (Figure 3.1). Based on previous protein-engineering CBP80 was expressed as CBP80ΔNLS lacking the first 19 amino acids including the NLS or as CBP80ΔNLSΔCC lacking additionally the coiled coil region (amino acids 653–701 replaced by a glycine) [138, 139]. CBP80ΔNLS was used for interaction studies and is in the following referred to as CBP80 and in complex with CBP20 as CBC. For crystallisation trials CBP80ΔNLSΔCC was additionally used.

These protein constructs served as basis for initial *in vitro* interaction studies before the interacting regions were narrowed down using shorter constructs. Figure 3.2 shows an example of the purified proteins. While most of the proteins were quite stable after purification, ARS2 and NCBP3 showed slight degradation (Figure 3.2). In general the proteins were expressed as His-tagged protein and purified using Ni affinity chromatography followed by TEV cleavage and a second Ni affinity chromatography to separate the cleaved tag from the protein. To remove any DNA or RNA contamination ion exchange chromatography was used. As final step and to check the quality of the protein the sample was subjected to SEC.

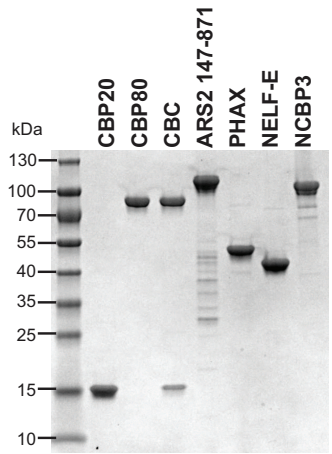


Figure 3.2 – Purified proteins

Coomassie-stained SDS-PAGE of purified proteins. ARS2 and NCBP3 showed slight degradation bands, whereas all other proteins are more stable and show only a single band.

3.1.2 *Biochemical Characterisation of the CBC-ARS2 Complex*

A direct interaction between ARS2$^{147\text{-}871}$ and CBC or the individual subunits CBP80 or CBP20 could not be observed by SEC and GST pull down (Figure 3.3A, B, data for subunits not shown). However, as previously described by Hallais *et al.* [73], the proteins co-eluted when CBC was immobilised on m^{7}GTP-Sepharose before incubation with ARS2 (data not shown). To test if the interaction is dependent on the cap analogue the SEC and GST pull down were repeated in the presence of m^{7}GTP, which resulted in the co-elution of CBC and ARS2 (Figure 3.3A, B). The result that a stable CBC-ARS2 complex was only formed in the presence of m^{7}GTP indicates a m^{7}GTP-dependent complex formation. This result was confirmed by ITC showing a twelve times higher affinity of ARS2 towards CBC in the presence of m^{7}GTP (K_D~1 µM compared to K_D~12 µM without m^{7}GTP) (Figure 3.3C). For ITC experiments C-terminal ARS2 constructs (ARS2$^{763\text{-}871}$ and ARS2$^{827\text{-}871}$) were used as longer versions tend to precipitate at the required concentrations.

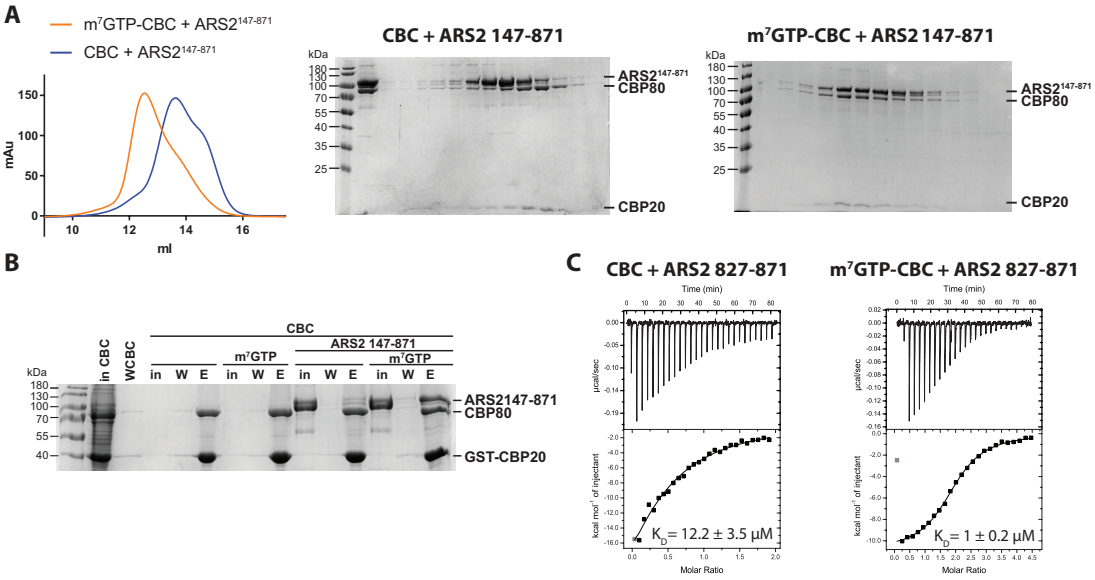


Figure 3.3 – Interaction of CBC-ARS2 is cap-dependent

A: SEC of CBC and ARS2 with and without m^{7}GTP. CBC and ARS2$^{147\text{-}871}$ were mixed with or without m^{7}GTP and subjected to gel filtration. Their elution profiles were overlaid and single fractions analysed by SDS-PAGE followed by Coomassie staining.
B: GST-CBP20/CBP80-ARS2 pull down in the absence and presence of m^{7}GTP. CBC containing GST-tagged CBP20 was immobilised on Glutathione resin and incubated with ARS2$^{147\text{-}871}$ in the presence and absence of m^{7}GTP. Input (in), last wash (W) and elution (E) fractions were analysed by Coomassie-stained SDS-PAGE.
C: Determination of the affinity of ARS2$^{827\text{-}871}$ for CBC in the presence and absence of m^{7}GTP. CBC or m^{7}GTP-CBC in the sample cell was titrated by ARS2$^{827\text{-}871}$. In each panel, the upper graph shows the raw data and the bottom graph shows the ligand concentration dependence of the heat released upon binding after normalisation. K_D values and standard deviation represent the average from at least two independent experiments.

Since the ITC data suggested different binding ratios (n=0.4-2) the complex was analysed by SEC-multi angle light scattering (MALS), which revealed an 1:1 stoichiometry in the presence of m^{7}GTP (see Appendix B.1 Figure B.1A). Thus, the observed ratio within the ITC experiment is likely due to inaccurate determination of the sample concentrations.

Using different ARS2 constructs the interacting region of ARS2 with CBC was narrowed down to the highly conserved C-terminal region of ARS2, compromising the last 27 amino acids (Figure 3.4A). A synthesised peptide containing ARS2$^{845\text{-}871}$ bound with the same affinity (K_D~1 µM) to m^{7}GTP-CBC as the longer constructs ARS2$^{763\text{-}871}$ and ARS2$^{827\text{-}871}$, while a construct missing the last 27 amino acids (ARS2$^{763\text{-}845}$) did not interact with CBC (Figure 3.4B). Thus, the highly conserved C-terminus of ARS2 (ARS2$^{845\text{-}871}$) is necessary and sufficient to bind to CBC (Figure 3.4). An overview of the analysed CBC-ARS2 interaction is given in Appendix B.1 Table B.1.

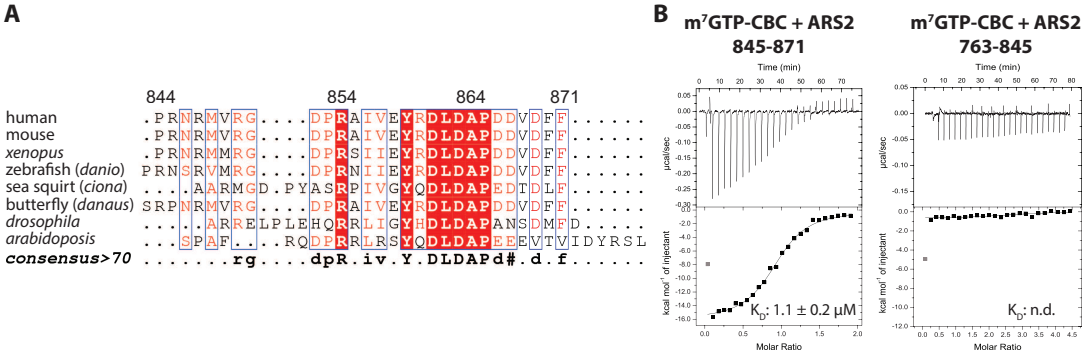


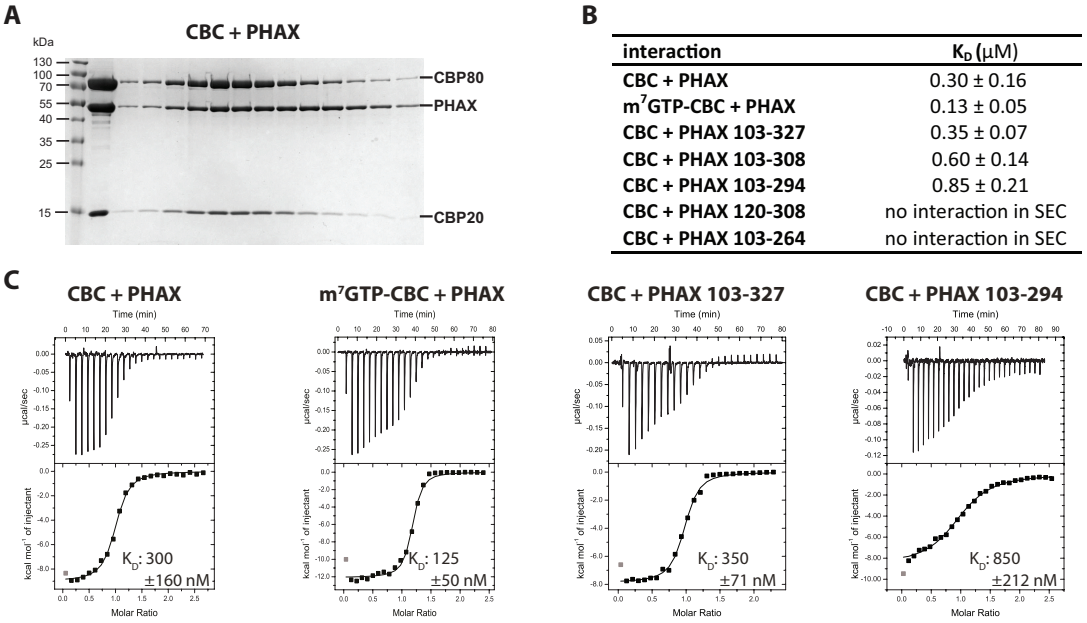
Figure 3.4 – The C-terminus of ARS2 mediates the interaction with CBC

A: Multiple sequence alignment of the C-terminus of ARS2. Conserved residues are highlighted in red and residues with a conservation above 70% are indicated below.
B: ITC of C-terminal ARS2 constructs with CBC in the presence of m^{7}GTP. m^{7}GTP-CBC in the sample cell was titrated by ARS2$^{763\text{-}845}$ or ARS2$^{845\text{-}871}$. The data were plotted as described in the caption to Figure 3.3C.

3.1.3 *Biochemical Characterisation of the CBC-PHAX Complex*

The CBC-PHAX complex, which has been reconstituted previously [158], was immediately obtained after mixing the two proteins and subjecting them to SEC (Figure 3.5A). However, no complex between CBP20 nor CBP80 alone and PHAX could be observed (data not shown). To determine the dissociation constant ITC experiments were performed, showing a K_D of 300 nM for CBC-PHAX in the absence and a K_D of 125 nM in the presence of the cap analogue m^{7}GTP (Figure 3.5C). Thus, the interaction of CBC and PHAX shows a slight, but not significant stabilisation of the CBC-PHAX complex upon cap-binding unlike observed previously for CBC-ARS2 (see Sec. 3.1.2). Using different constructs for SEC and ITC the binding site of PHAX on CBC could be narrowed down to the residues 103-294 (Figure 3.5B), containing

most of the RRM domain of PHAX (residues 227-308). By ITC a K_D of 350 nM was measured for binding of PHAX$^{103-327}$ to CBC in the absence of m^{7}GTP, comparable to full length PHAX, whereas PHAX$^{103-308}$ and PHAX$^{103-294}$ showed slightly reduced affinities with K_Ds of 600 and 850 nM, respectively (Figure 3.5C). The analysed CBC-PHAX interactions are summarised in Figure 3.5B.



interaction	K_D (µM)
CBC + PHAX	0.30 ± 0.16
m^{7}GTP-CBC + PHAX	0.13 ± 0.05
CBC + PHAX 103-327	0.35 ± 0.07
CBC + PHAX 103-308	0.60 ± 0.14
CBC + PHAX 103-294	0.85 ± 0.21
CBC + PHAX 120-308	no interaction in SEC
CBC + PHAX 103-264	no interaction in SEC

Figure 3.5 – PHAX binding to CBC

A: Reconstitution of the CBC-PHAX complex. CBC was incubated with molar excess of PHAX before subjected to SEC. Single fractions were analysed by SDS-PGAE followed by Coomassie-staining.
B: Overview of PHAX-CBC binding studies. Different PHAX constructs were tested for CBC binding using SEC and ITC. In case of co-elution from the SEC the K_D was determined using ITC.
C: ITC data and curve fit to determine the affinity of PHAX for CBC in the presence and absence of m^{7}GTP. CBC or m^{7}GTP-CBC in the sample cell was titrated by PHAX. The data were plotted as described in the caption to Figure 3.3C.

3.1.4 *Biochemical Characterisation of the CBC-NELF-E Complex*

The interaction of CBC with the N-terminal part of NELF-E$^{244-380}$ has been described previously [156] and therefore this construct, which could be expressed in *E. coli*, was used along with the full length protein expressed in insect cells for interaction studies. When analysing the binding affinities for CBC and NELF-E$^{244-380}$ by ITC a cap-dependent 8-fold increase of the dissociation constant was detected (K_D~0.4 µM to K_D~0.05 µM) (Figure 3.6D). Consistent with this and similar to ARS2 (see Sec. 3.1.2), a stable CBC-NELF-E complex was only formed in the presence of the cap analogue m^{7}GTP (Figure 3.6A). Already in the absence of m^{7}GTP CBC and NELF-E

showed slight co-elution (Figure 3.6A middle), indicating a weak interaction of the two proteins, which could be stabilised by the addition of m^{7}GTP (Figure 3.6A right). Additionally, as observed before by Narita *et al.* [156], the deletion of the last 20 amino acids (NELF-E$^{244\text{-}360}$) abolished CBC binding completely, since no interaction between NELF-E$^{244\text{-}360}$ and m^{7}GTP-CBC could be detected by either SEC or ITC (Figure 3.6B, Appendix B.3 Table B.3). To test whether the highly conserved C-terminal 21 amino acids (Figure 3.6C) are sufficient for CBC binding, the synthetic peptide NELF-E$^{360\text{-}380}$ was used for ITC binding studies. With a K_D of 3.3 µM NELF-E$^{360\text{-}380}$ showed a much weaker affinity for m^{7}GTP-CBC than NELF-E$^{244\text{-}380}$ (K_D~0.05 µM) (Figure 3.6D). Extending the peptide to 27 amino acids NELF-E$^{354\text{-}380}$ resulted in the

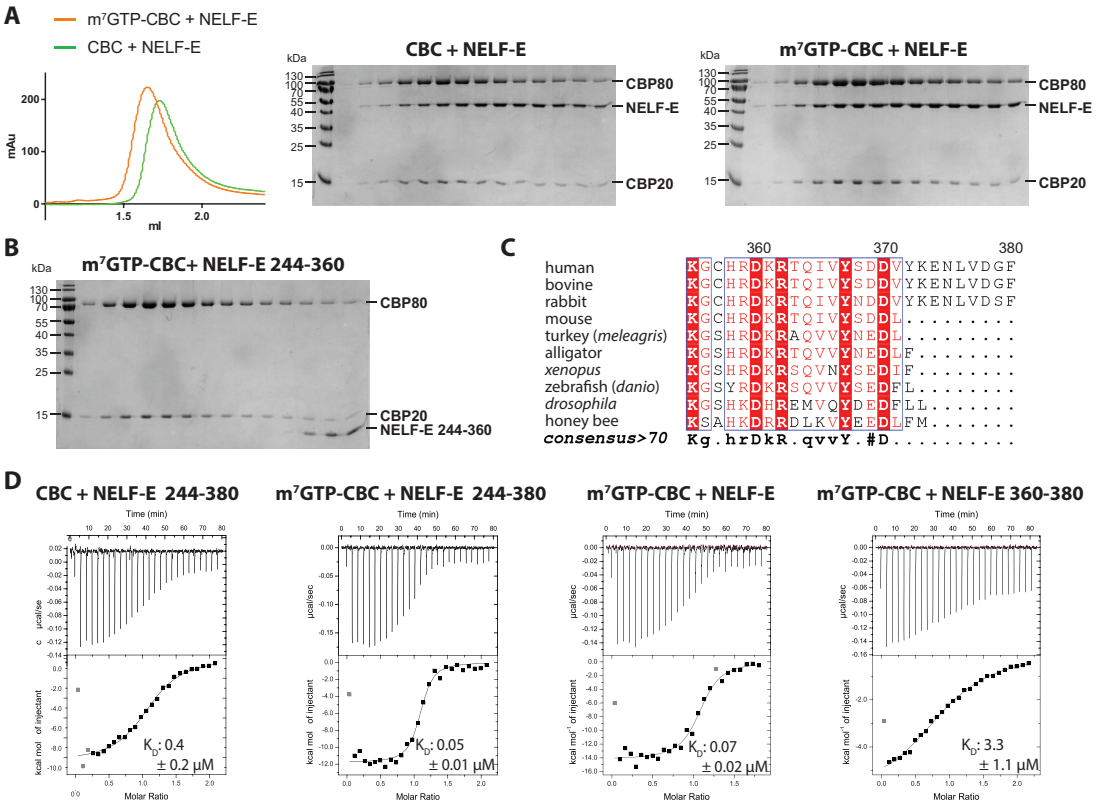


Figure 3.6 – NELF-E C-terminus binds CBC in a cap-dependent manner

A: Interaction of NELF-E and CBC in absence and presence of m^{7}GTP. CBC and NELF-E were mixed with or without m^{7}GTP and subjected to gel filtration. Their elution profiles were overlaid and single fractions analysed by Coomassie-stained SDS-PAGE.
B: SEC of CBC and NELF-E$^{244\text{-}360}$. CBC was mixed with molar excess of NELF-E$^{244\text{-}360}$and m^{7}GTP and subjected to SEC. Protein containing fractions were analysed by Coomassie stained SDS-PAGE.
C: Sequence alignment of the C-terminus of NELF-E. Conserved residues are highlighted in red and residues with a conservation above 70% are indicated below.
D: Affinity determination for NELF-E-CBC binding in the presence and absence of m^{7}GTP. CBC or m^{7}GTP-CBC in the sample cell was titrated by different NELF-E constructs. The data were plotted as described in the caption to Figure 3.3C.

same weak affinity (K_D~3.8 µM) (see Appendix B.3 Table B.3). This suggests that the C-terminus of NELF-E plays an important part in the interaction with CBC, though the interaction is also stabilised by further protein-protein interactions. Notably, these interactions are not sufficient to obtain the CBC-NELF-E complex in the absence of C-terminal 20 amino acids of NELF-E (NELF-E$^{244\text{-}360}$). A summary of the studied CBC-NELF-E interactions is given in Appendix B.3 Table B.3.

3.1.5 *Compatibility of PHAX, ARS2 and NELF-E Binding to CBC*

After identifying the minimum CBC binding regions of ARS2, PHAX and NELF-E the next question was, if and which proteins can bind CBC at the same time *in vitro*. Earlier it has been shown that ARS2 binding is compatible with PHAX binding [73]. A stable complex containing CBC, PHAX and ARS2$^{147\text{-}871}$ could be reconstituted by incubating the proteins together with m^{7}GTP and subjecting them to SEC (Figure 3.7A). However, no complex containing CBC-NELF-E and PHAX nor ARS2 was observed using SEC (see Appendix B.4 Figure B.7). To further analyse if the interactions of NELF-E and ARS2 with CBC are compatible, preformed CBC-NELF-E com-

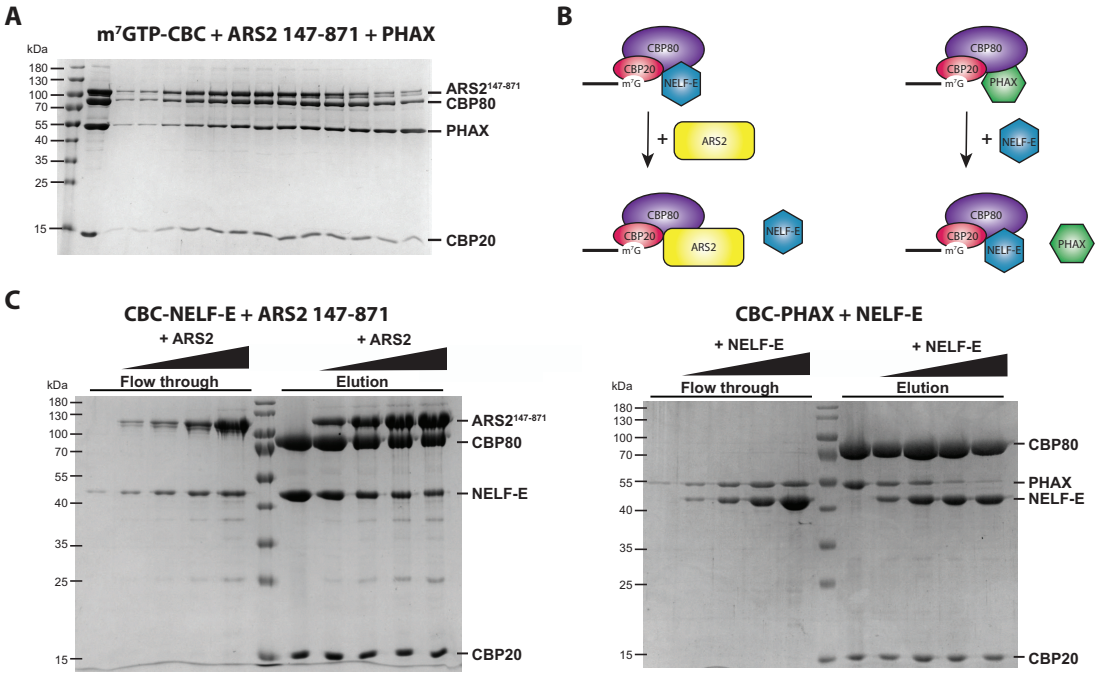


Figure 3.7 – Competition between NELF-E and ARS2/PHAX for CBC binding

A: Reconstitution of the m^{7}GTP-CBC-ARS2-PHAX complex. ARS2$^{147\text{-}871}$ was incubated with molar excess of m^{7}GTP-CBC and PHAX before subjected to SEC. The protein containing fraction were analysed by Coomassie-stained SDS-PAGE.
B: Set up of m^{7}GTP-Sepharose displacement pull down. CBC-NELF-E or CBC-PHAX were immobilised on m^{7}GTP-Sepharose and incubated with different amounts of ARS2$^{147\text{-}871}$ or NELF-E, respectively. Flow through and elutions were analysed by Coomassie-stained SDS-PAGE (**C**).

plex was immobilised on m^{7}GTP-Sepharose and incubated with different amounts of ARS2$^{147\text{-}871}$ (Figure 3.7B left). The protein bound to the resin and the flow through for each sample were then analysed by SDS-PAGE. The results showed that with increasing amounts of ARS2$^{147\text{-}871}$ more NELF-E was in the flow through and less NELF-E remained bound to CBC, whereas, more ARS2 co-eluted with CBC (Figure 3.7C left). Thus, ARS2 was competing with NELF-E for CBC binding. Performing the same assay with preformed CBC-PHAX and adding NELF-E to the immobilised complex (Figure 3.7B right) resulted in the displacement of PHAX from CBC (Figure 3.7C right), showing that also PHAX and NELF-E are competing for CBC binding. Titrating ARS2$^{147\text{-}871}$ to a preformed CBC-PHAX within the same experimental set up resulted in a stable complex containing CBC, ARS2 and PHAX (data not shown).

The same competition was observed in a fluorescence polarisation assay using N-terminal FAM-labeled ARS2$^{845\text{-}871}$ (ARS2$^{845\text{-}871}$(FAM)) or FAM-labeled NELF-E$^{354\text{-}380}$ (NELF-E$^{354\text{-}380}$(FAM)) (Figure 3.8). As expected, addition of m^{7}GTP-CBC to either of these labeled peptides led to an increase in fluorescence polarisation, indicating binding (Figure 3.8A, D). Further addition of NELF-E$^{244\text{-}380}$ or NELF-E$^{360\text{-}380}$ to the preformed m^{7}GTP-CBC-ARS2$^{845\text{-}871}$(FAM) complex led to dissociation of CBC from ARS2$^{845\text{-}871}$(FAM) shown by a decreasing polarisation (Figure 3.8B, data not shown for NELF-E$^{360\text{-}380}$). The same result was also obtained for the addition of ARS2$^{827\text{-}871}$ or ARS2$^{845\text{-}871}$ to the preformed m^{7}GTP-CBC-NELF-E$^{354\text{-}380}$(FAM) complex (Figure 3.8E, data not shown for ARS2$^{845\text{-}871}$). Thus, ARS2 and NELF-E were competing for CBC binding. However, titrating PHAX$^{103\text{-}327}$, that shows a better behaviour than the full length protein at high protein concentration, to m^{7}GTP-CBC-ARS2$^{845\text{-}871}$(FAM) or m^{7}GTP-CBC-NELF-E$^{354\text{-}380}$(FAM) led to no change in polarisation (Figure 3.8F, G). Interestingly, the addition of PHAX$^{103\text{-}327}$ to a mixture of CBC, FAM-ARS2$^{827\text{-}871}$ and NELF-E$^{244\text{-}380}$ showed an increase in polarisation, indicating that PHAX can replace NELF-E$^{244\text{-}380}$ from CBC and allows ARS2$^{845\text{-}871}$(FAM) to bind again CBC (Figure 3.8C). This demonstrates that binding of NELF-E$^{354\text{-}380}$ but not NELF-E$^{244\text{-}380}$ to CBC is compatible with PHAX$^{103\text{-}327}$ binding, suggesting that NELF-E locates at the same position around CBC as PHAX. Notably, for both proteins these competing domains include the RRM domains, which are the predicted most structures parts of PHAX and NELF-E (see Figure 3.1). However, for ARS2 and NELF-E the C-terminal peptides were sufficient to compete with each other, respectively. To exclude that the changes in polarisation were due to interactions of NELF-E or PHAX with ARS2$^{845\text{-}871}$(FAM) and ARS2 or PHAX with NELF-E$^{354\text{-}380}$(FAM), the same experiments were performed in the absence of CBC. All these control experiments showed no significant change in polarisation (see Appendix B.4 Figure B.8), indicating that the observed changes in polarisation are caused by interactions of the labeled peptide with CBC.

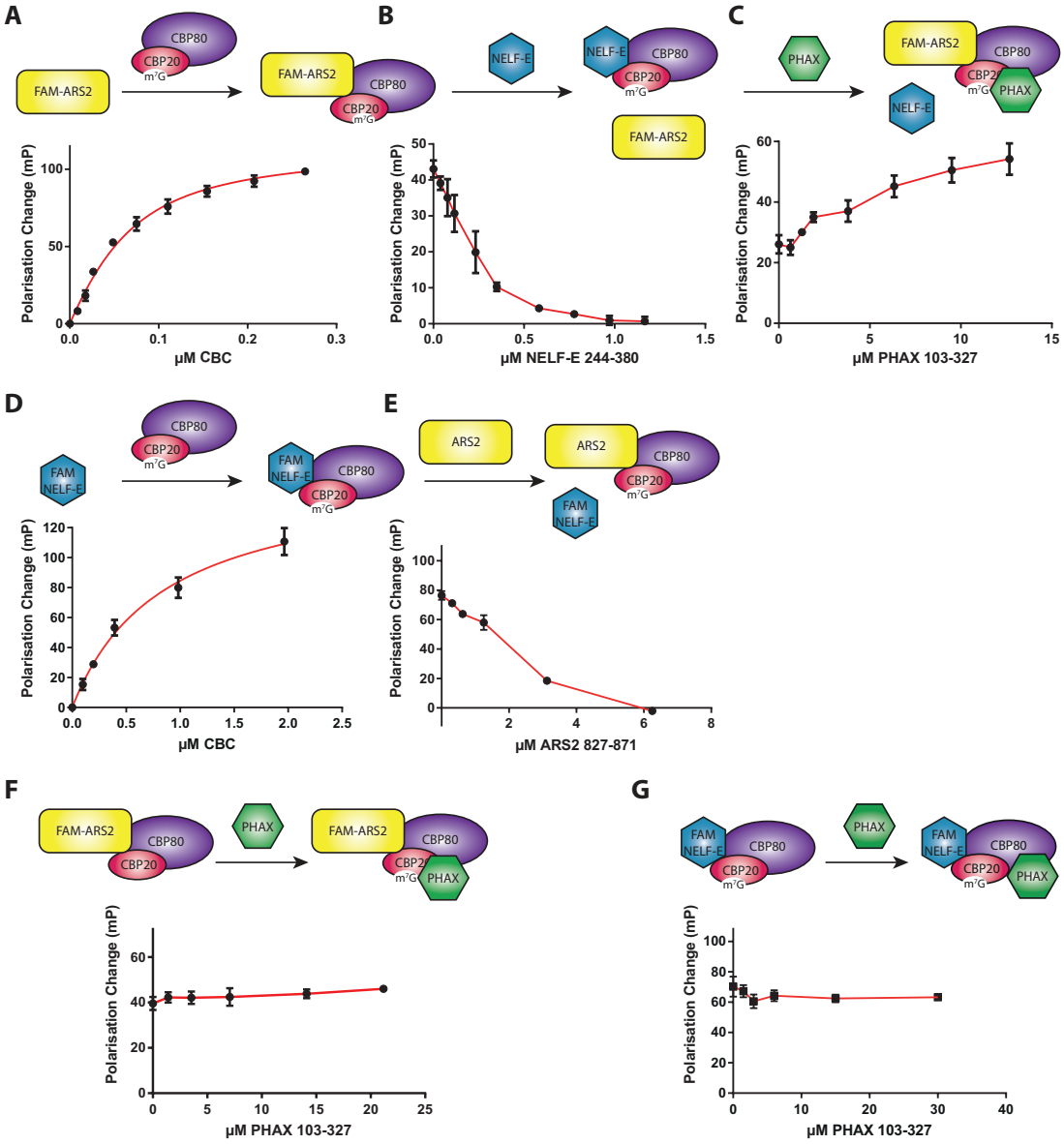


Figure 3.8 – Fluorescence polarisation assay

Fluorescence polarisation data showing the titration of
A: m^{7}GTP-CBC to ARS2$^{845\text{-}871}$(FAM).
B: NELF-E$^{244\text{-}380}$ to m^{7}GTP-CBC-ARS2$^{845\text{-}871}$(FAM).
C: PHAX$^{103\text{-}327}$ to m^{7}GTP-CBC-ARS2$^{845\text{-}871}$(FAM)/NELF-E$^{244\text{-}380}$.
D: m^{7}GTP-CBC to NELF-E$^{354\text{-}380}$(FAM).
E: ARS2$^{827\text{-}871}$ to m^{7}GTP-CBC-NELF-E$^{354\text{-}380}$(FAM).
F: PHAX$^{103\text{-}327}$ to m^{7}GTP-CBC- ARS2$^{845\text{-}871}$(FAM).
G: PHAX$^{103\text{-}327}$ to m^{7}GTP-CBC-NELF-E$^{354\text{-}380}$(FAM).
Displayed are the median and error bars (SD) of three independent experiments.

3.2 STRUCTURAL CHARACTERISATION OF CBC COMPLEXES

3.2.1 *Structure of m^{7}GTP-CBC-ARS2*

3.2.1.1 *Structure Determination of m^{7}GTP-CBC-ARS2*

To investigate the structural basis for the CBC-ARS2 interaction, crystallisation trials of CBCΔNLS and CBCΔNLSΔCC with different CBC binding ARS2 constructs (see Appendix B.1 Table B.1) were performed in the presence of m^{7}GTP or m^{7}GpppG. Only the complex containing m^{7}GTP-CBCΔNLS-ARS2$^{827\text{-}871}$ purified by SEC resulted in three different crystal forms after 24-72 h at 20°C (not shown). Two of these crystals could be manually reproduced and the conditions were further optimised to obtain bigger monocrystalline crystals. Most of the crystals showed only weak diffraction or no extra density for the peptide. The best crystal showing unambiguous density for the ARS2 peptide diffracted to 2.8 Å resolution. Initial phases were obtained by molecular replacement using the known m^{7}GpppG-CBC complex structure (PDB:1H2T [139]) as search model. The crystal belonged to space group *P1* with cell dimensions of α=γ=90°, β=90.3° and a=70.52, b=122.99 and c=270.98. There are eight complexes in the asymmetric unit with variable occupancy for the ARS2 peptide. Data collection and refinement statistics are summarised in Table 3.1 and Table 3.2. The omit difference density map of the ARS2 peptide is shown in Figure 3.9C.

Table 3.1 – Data collection statistics for m^{7}GTP-CBCΔNLS-ARS2$^{827\text{-}871}$

CRYSTAL	m^{7}GTP-CBCΔNLS-ARS2$^{827\text{-}871}$
Beamline	ID23-1
Wavelength (Å)	0.97917
Space group	*P*1
Cell dimensions (Å)	a=70.52 b=112.99 c=270.98
Cell angles (°)	α=γ=90 β=90.30
Resolution range of data (last shell) (Å)	50.0-2.80 (2.80-2.80)
Completeness (last shell) (%)	97.5 (96.3)
R-sym (last shell) (%)	9.7 (69.7)
I/σI (last shell)	6.63 (1.20)
CC(1/2) (last shell)	0.995 (0.585)
Redundancy (last shell)	2.08 (2.00)

Table 3.2 – Refinement statistics for m^{7}GTP-CBCΔNLS-ARS2$^{827\text{-}871}$

CRYSTAL	m^{7}GTP-CBCΔNLS-ARS2$^{827\text{-}871}$
Reflections used in refinement work (free)	190778 (10059)
R-work (last shell)	0.230 (0.382)
R-free (last shell)	0.266 (0.364)
Number of non-hydrogen atoms	58818
Protein	57540 (8x CBC)
Ligand (m^{7}GTP)	264 (8x m^{7}GTP)
Peptide	938 (chains C, F, I, L, O, R, U, X)
Geometry and B-factors	
RMSD (bonds)	0.008
RMSD (angles)	1.194
Ramachandran favored (%)	97.9
Ramachandran outliers (%)	0.03
Clash score	0.86
MolProbity score	0.88
Average B-factor	73.3
Protein	72.8 (8x CBC)
Ligand (m^{7}GTP)	66.3 (8x m^{7}GTP)
Peptide	106.1 (chains C, F, I, L, O, R, U, X)

3.2.1.2 *Structure Analysis of m^{7}GTP-CBC-ARS2*

The overall structure of CBC and the binding of CBP20 to the cap analogue in the ARS2 bound complex is similar to the cap bound CBC [139]. Only locally ARS2 perturbs the interaction of CBP20 with CBP80 (Figure 3.9A). The ARS2 peptide, which is visible from residues 850-861 and 868-871, forms an extended chain, which wraps around the CBP80-CBP20 interface, interacting with both subunits (Figure 3.9A). There are two main binding regions of the ARS2 peptide: one involving residues 852-861 (proximal site) and the other one involving the extreme C-terminal residues 868-871 (distal site). For the highly conserved residues 862-867, which are linking the interacting region, no electron density is present. The main interacting regions are residues 460-463, 557-570, 607-616 and 647-651 of CBP80 and residues 48-55 and 71-80 of CBP20. These CBP20 residues belong to the RNP domain, namely the β1-α2 loop and the β2-β3 loop. The total buried surface area upon ARS2 peptide binding to CBC is 2257 Å^2 (PISA server [108]) shared 52% and 48% respectively between CBP80 and CBP20.

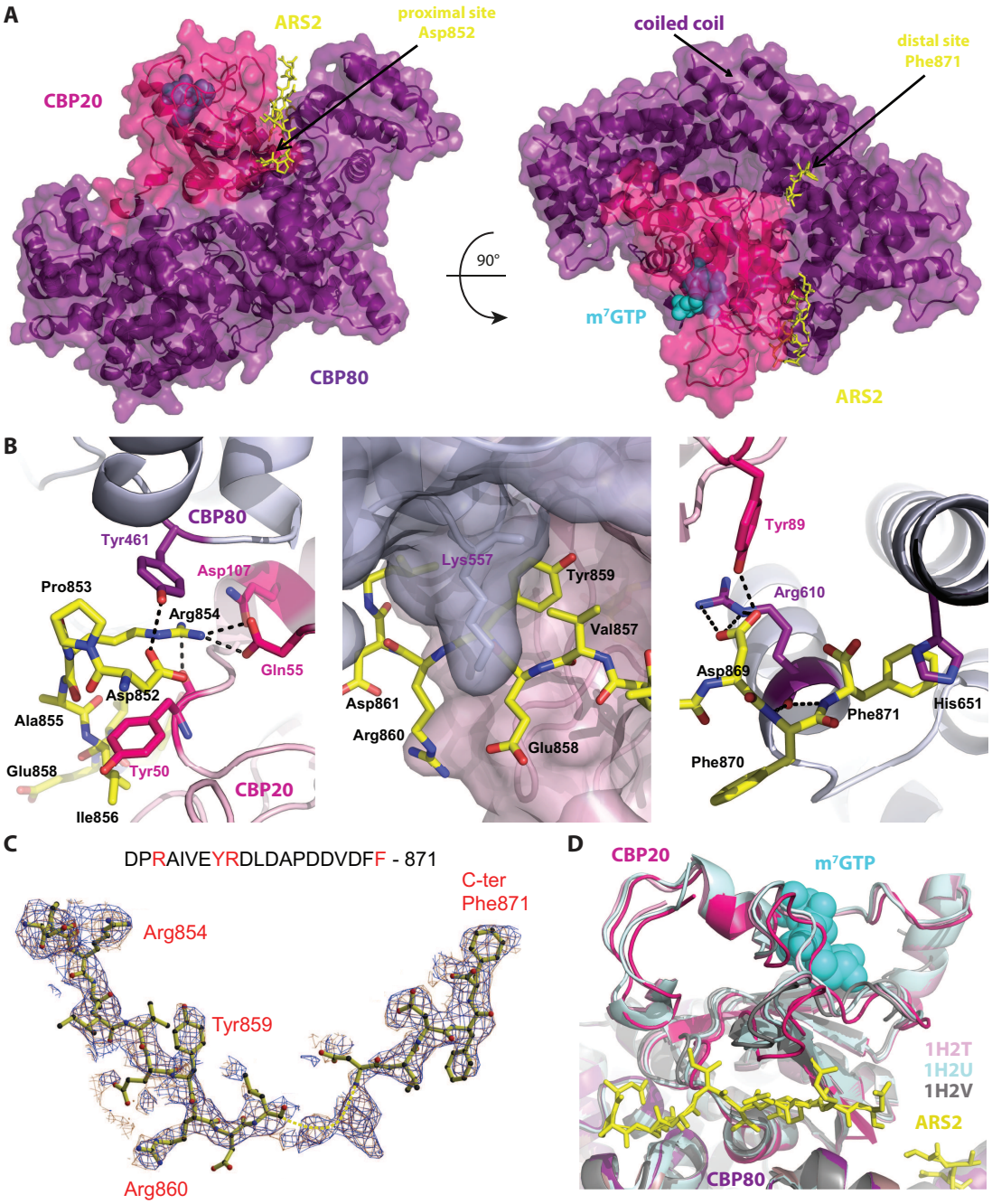


Figure 3.9 – Crystal structure of m^{7}GTP-CBCΔNLS-ARS2$^{827\text{-}871}$

A: Overview of the m^{7}GTP-CBCΔNLS-ARS2$^{827\text{-}871}$ crystal structure. The C-terminus of ARS2 (yellow) lies in the interface of CBP80 (purple) and CBP20 (pink), which binds m^{7}GTP (cyan). **B:** Close-up view of the key interacting residues. Black dashed lines indicate polar contacts. **C:** Omit difference density map of the ARS2 peptide contoured at 2σ. The main interacting residues are labeled and the highly conserved residues within the C-terminus of ARS2 are highlighted in red within the sequence.
D: Superposition of CBC structures (1H2V apo (gray) and 1H2T (rose) and 1H2U (light blue) [139] bound to m^{7}GpppG (cyan)) with m^{7}GTP-CBC-ARS2$^{827\text{-}871}$ via CBP80 shows successive relative displacements of the CBP20 subunit.

In the proximal binding site including ARS2 residues 852-861 Asp852, Arg854 and Tyr859 are the main interacting residues (Figure 3.9B, left). Asp852 stacks on Tyr50 of CBP20 and electrostatically stabilises Arg854, which stacks via its guanidinium group on Tyr461 of CBP80 and forms a salt bridge with Asp107 from CBP20. In the crystal structure without the ARS2 peptide, the hydroxyl groups from CBP80 Tyr461 and CBP20 Tyr50 interact at the intersubunit interface [139]. However, this conformation is sterically incompatible with the presence of ARS2 Asp852 and Arg854. Thus, in the presence of the ARS2 peptide CBP20 Tyr50 adopts an alternative conformation allowing to interact with ARS2 Asp852. Additionally, Ile856 makes hydrophobic contacts with CBP20 Tyr50 and Thr79. Further along the peptide chain ARS2 Tyr859 is inserted between CBP20 and CBP80 (Figure 3.9B, middle) making contacts with CBP20 Met71 and Leu73 and CBP80 Ser558 and His561. Furthermore, its hydroxyl group points towards the main-chain amino group of CBP20 Glu53, whose side-chain makes CBC intersubunit interactions. Superposition of CBP80 of the m^7GpppG-CBC complexes (RMSD: 0.35 Å for 1H2T and 0.38 Å for 1H2U [139]) on the m^7GTP-CBC-ARS2 complex shows that the CBP20 β2-β3 loop (residues 71-81) and C-terminal extension of the RRM domain (residues 126-150) are slightly shifted (Figure 3.9D). This results in the closure of the groove in which the ARS2 peptide binds, with a maximum displacement of 2.0 Å. Moreover, it allows the main-chain of CBP20 residues 78-80 to run anti-parallel to the ARS2 residues 857-859. There are hydrogen bonds between the carbonyl oxygen of Lys78 with the amino group of Tyr859 and the carbonyl group of Val857 with the amino group of Ala80 . Additional contacts are made between Arg860 and CBP20 Leu73 (Appendix B.1 Figure B.2).

In the distal site the terminal ARS2 residue Phe871 is buried in a hydrophobic pocket at the base of the long, solvent exposed coiled-coil of CBP80, formed by CBP80 Ile609, Cys616, Met648 and His651. The side-chain of His651 partially stacks on Phe871 and also interacts electrostatically with the C-terminal carboxylate of ARS2 (Figure 3.9B, right). Moreover, CBP20 residue Tyr89 mediates hydrogen bonds with ARS2 Asp869 and stacks on CBP80 Arg610, whose carbonyl-oxygen hydrogen interacts with the amino groups of ARS2 Phe870 and Phe871.

All of the cited interacting residues are highly conserved in metazoan ARS2 (Figure 3.4A) or CBC suggesting conservation of this interaction and underscoring its biological importance. An overview of all interactions can be found in Appendix B.1 Figure B.2A.

In light of the structural organisation of the CBC-ARS2 complex, the increased affinity of ARS2 to cap bound CBC can be explained by comparing the crystal structure of apo-CBC (PDB:1H2V [139]) (Figure 3.9D). Upon binding of the cap analogue m^7GTP CBP20 acquires a more structured N- and C-terminal region, which also results in an initial closure of the groove between CBP20 and CBP80, in which the ARS2 peptide binds (Figure 3.9D). This probably leads to a stabilisation of the interaction between CBC and the ARS2 peptide, which is important for the CBC-ARS2 interaction *in vitro*.

3.2.2 *Validation of the Observed CBC-ARS2 Interaction Site*

To validate the observed binding site, point mutations were introduced into ARS2 as well as in the corresponding interacting residues of CBP80 and CBP20 using site directed mutagenesis. These mutants were then used for subsequent binding studies.

For ARS2 the in *E. coli* expressed and well behaved ARS2$^{827\text{-}871}$ construct was mutated and used for ITC experiments in the presence of the cap analogue m^{7}GTP (Figure 3.10A). The mutation of F871D led to a slight decrease in affinity (K_D~1.8 µM) compared to the wild type construct (K_D~1 µM, see Figure 3.3C). However, the double mutation of R854A and Y859A weakened the interaction considerably, so that no K_D could be obtained from the binding data. Introducing all three point mutations together completely abolished the interaction with CBC. This data show that R854 and Y859 are essential for the interaction between CBC and ARS2, whereas F871 plays a minor role.

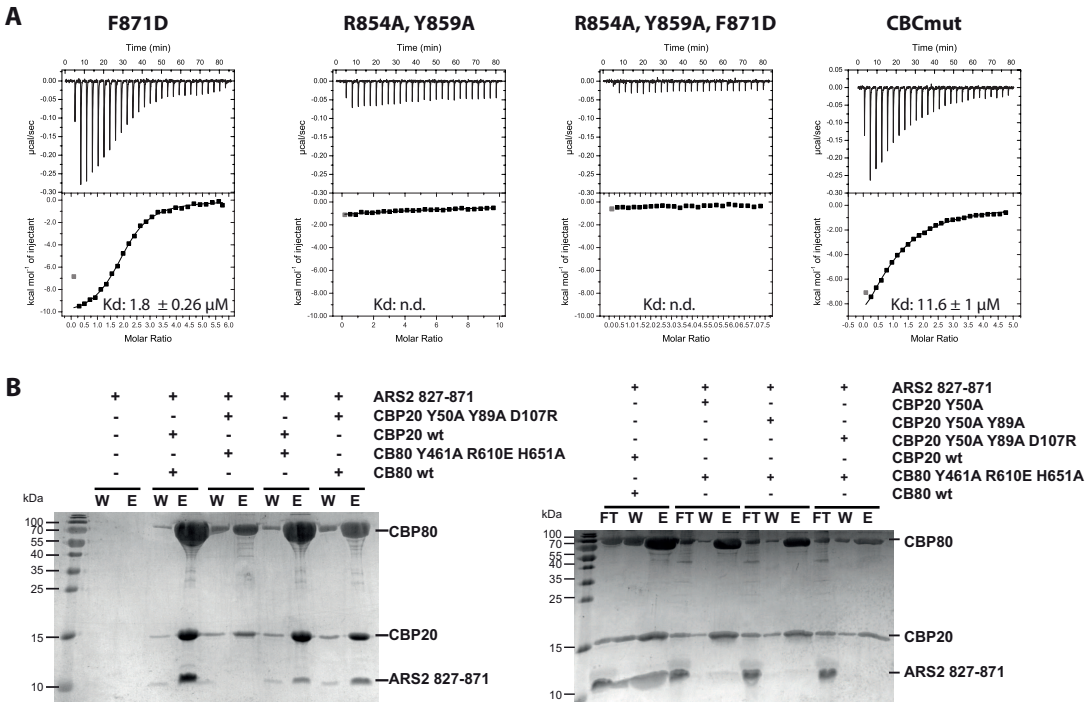


Figure 3.10 – Abolishing the interaction by mutating ARS2 and CBC

A: Analysis of ARS2 mutants for CBC binding. Mutations were introduced into the main interacting residues of ARS2 and their effects on CBC binding were tested by ITC. ARS2$^{827\text{-}871}$was titrated to CBC in the presence of m^{7}GTP. All mutations led to a lower K_D compared to wild type ARS2 827-871 (K_D~1 µM, see Sec. 3.1.2). The data were plotted as described in the caption to Figure 3.3C.

B: Analysis of CBC mutants for ARS2 binding. Mutations were introduced into the main interacting residues of CBP80 and CBP20. The CBC complexes were reconstituted and immobilised on m^{7}GTP-Sepharose. After incubation with ARS2$^{827\text{-}871}$ the wash (W) and elution (E) fractions were analysed via SDS-PAGE followed by Coomassie staining. The right gel shows additionally the flow through (FT).

For CBC mutations in CBP80 Y461A R610E H651A as well as CBP20 Y50A Y89A D107R were introduced and initially tested for ARS2 binding via pull down using m^{7}GTP-Sepharose. As CBP80 was expressed in insect cells all mutations were introduced at once, whereas for CBP20 expressed in *E. coli* mutations were introduced one after the other resulting in CBP20 Y50A, CBP20 Y50A Y89A and CBP20 Y50A Y89A D107R mutants. The immobilised CBP20/CBP80 complexes were incubated with ARS2$^{827\text{-}871}$ and the elutions analysed by SDS-PAGE. The obtained results (Figure 3.10B) showed that mutations in all three residues in CBP20 or CBP80, respectively, are not sufficient to abolish ARS2 binding. Also the combination of CBP80 Y461A R610E H651A with CBP20 Y50A or CBP20 Y50A Y89A still showed slight affinity for ARS2. Only the combination of mutations in all three residues in CBP20 and CBP80 led to no detectable co-elution of ARS2 . However, compared with the other CBC elutions less of the complex carrying the six mutations eluted from the resin, indicating reduced affinity for the cap analogue m^{7}GTP.

Next ITC experiments were performed to determine the affinity of the mutated CBC (Table 3.3). To exclude side effects caused by the introduced mutations, the binding of m^{7}GTP and PHAX were also tested to ensure that the overall conformation of CBC did not change upon mutation and only the interaction with ARS2 is abolished. The triple mutant CBP20 Y50A Y89A D107R resulted in a reduced m^{7}GTP affinity (K_D~70 µM compared to wild type K_D~0.1 µM) and was therefore excluded from further experiments. CBC consisting of CBP80 Y461A R610E H651A and CBP20 Y50A or CBP20 Y50A Y89A were binding m^{7}GTP and PHAX with the same affinity

Table 3.3 – Interaction studies of CBC mutants

Using ITC the interactions of different CBC constructs were tested and compared to the wild type.

INTERACTION	K_D
m^{7}GTP + CBC	0.112 ± 0.045 µM
m^{7}GTP + CBC (CBP80 Y461A R610E H651A + CBP20 Y50A Y89A D107R)	70 ± 14 µM
m^{7}GTP + CBC (CBP80 Y461A R610E H651A + CBP20 Y50A Y89A)	0.148 ± 0.025 µM
m^{7}GTP + CBCmut (CBP80 Y461A R610E H651A + CBP20 Y50A)	0.095 ± 0.024 µM
CBC (CBP80 Y461A R610E H651A + CBP20 Y50A Y89A) + PHAX	0.6 ± 0.23 µM
CBCmut (CBP80 Y461A R610E H651A + CBP20 Y50A) + PHAX	0.5 ± 0.14 µM
m^{7}GTP-CBC (CBP80 Y461A R610E H651A + CBP20 Y50A Y89A) + ARS2 827-871	10.9 ± 0.5 µM
m^{7}GTP-CBCmut (CBP80 Y461A R610E H651A + CBP20 Y50A) + ARS2 827-871	11.6 ± 1.1 µM

as CBC wild type and were therefore used for further ARS2 binding studies. In ITC experiments both CBC mutants showed an eleven to twelve times weaker affinity (K_D~11-12 µM) to $ARS2^{827-871}$ in the presence of m^7GTP (Table 3.3), confirming that the mutated residues are involved in ARS2 binding. As no additional change in affinity was observed for CBC containing the additional mutation Y89A in CBP20 compared to CBP20 Y50A, CBC carrying mutations in CBP20 Y50A and CBP80 Y461A R610E H651A was used for further experiments and is referred to as CBCmut.

To test the mutations in full length CBC and ARS2 the bridged yeast two-hybrid system was used. Here Alix (ALG-2-interacting protein X), a cytoplasmic protein involved in apoptosis, served as negative control. CBP80 was fused to the GAL4 DNA-binding domain (pAS2) and expressed together with CBP20 (p422) in Y187 strains, while ARS2 and Alix on pACTII fused to the GAL4 transcription activating domain were transformed into CG1945 strain. After crossing the strains and plating them on selection media, only colonies for CBC containing both wild type subunits and ARS2 wild type could grow on the quadruple-selective media (Table 3.4). Also a few colonies of ARS2 wild type crossed with CBP80 wild type CBP20 Y50A were obtained, showing that the mutation in CBP20 alone is not sufficient to abolish completely the interaction with ARS2. For the mutated ARS2 construct no colonies were obtained. This showed that mutation of all three ARS2 residues or mutations in both CBC subunits (CBP80 Y461A R610E H651A + CBP20 Y50A) were sufficient to abolish the interaction in the bridged yeast two-hybrid system (Table 3.4). Overall these results confirmed the ITC and pull down data and demonstrated additionally that the low affinity of the CBC mutant for ARS2 detected in the pull down and ITC experiments is not strong enough to form a stable CBC-ARS2 complex within the yeast two-hybrid system.

Table 3.4 – Bridge yeast two-hybrid assay using CBC and ARS2 mutants

CBP80 fused to the GAL4 DNA binding domain in pAS2 and CBP20 in p422 were transformed into Y187. ARS2 and the control Alix were transformed as GAL4 transcription activating domain fused proteins (pACTII) in CG1945. After mating of haploid strains, the diploid strains were selected on –Ala-Leu-Thr media and analysed for their interaction via their growth on –Ala-Leu-Thr-His media.
– indicates no colony, +/- few colonies and + many colonies.

pAS2 / pACT II	ARS2 wt	ARS2 R854A Y859A F871A	Alix (control)
CBP80 wt	-	-	-
CBP80 Y461A R610E H651A	-	-	-
CBP80 wt + CBP20 wt	+	-	-
CBP80 wt + CBP20 Y50A	+ / -	-	-
CBP80 Y461A R610E H651A + CBP20 Y50A	-	-	-
CBP20 Y50A	-	-	-

3.2.3 *Structural Characterisation of the CBC-PHAX Complex*

To better understand the interaction of CBC and PHAX, crystallisation trials containing CBC, different PHAX constructs and/or m^{7}GTP or RNA were set up, but no diffracting crystal could be obtained. Nevertheless, to map the PHAX binding site on CBC cross-linking mass spectrometry was performed using the two amine-reactive, homo-bifunctional reagent disuccinimidyl suberate (DSS). The cross-linked peptides were subjected to mass spectrometric analysis and stringently identified using the xQuest/xProphet software by the EMBL Proteomics Core Facility [117]. All identified intra-protein cross-links are listed in Appendix B.2 Table B.2. Although only a few cross-links could be identified, four different CBC-PHAX inter-protein cross-links were obtained: three between CBP20 and PHAX and one between CBP80 and PHAX (Figure 3.11A, B). The identified cross-linked lysines from PHAX K163, K173, K216 and K256 lie all within residues 103-294, which were identified to be essential for CBC binding (see Sec. 3.1.3). The cross-linked lysines of CBC are located close to

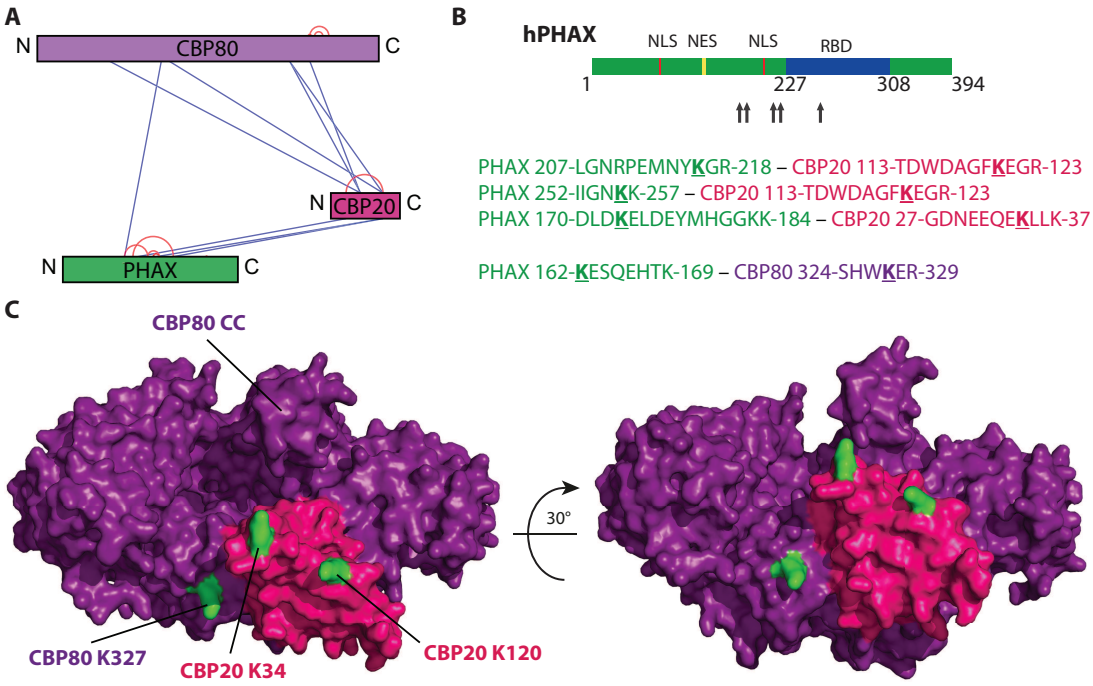


Figure 3.11 – Mapping of the CBC-PHAX interaction site

CBC-PHAX was cross-linked using the lysine-specific bifunctional cross-linker DSS, fragmented by proteolysis and cross-linked peptides were identified by mass spectrometry.
A: Overview of intra- (red) and inter-protein (blue) cross-links of CBC-PHAX.
B: Cross-links identified between CBP80 (purple) or CBP20 (pink) and PHAX (green). Location of the cross-linked lysines of PHAX are indicated as arrows on the depiction of PHAX. Identified cross-linked lysines of PHAX and CBP80 or CBP20 within the cross-linked peptide sequence are underlined and in bold.
C: Location of cross-linked CBC residues. Cross-linked lysine residues of CBC are highlighted in green within the CBC crystal structure (PDB:1H2V [139]) with CBP80 in purple and CBP20 in pink.

each other on the surface (Figure 3.11C). CBP20 K34 and K120 belong to the more unstructured region of CBP20 before and after its RNP domain. CBP80 K327 is part of the loop connecting helix α15 and α16 next to the CBP20 interface. All these residues are facing into the grove between CBP20 and the coiled coil of CBP80 next to the ARS2 binding site (see Figure 3.9). In line with the cap-independent binding of PHAX to CBC (see Sec. 3.1.3), this area does not undergo big conformational change upon cap-binding [139]. Furthermore, the data provide evidence that residues of CBP20 and CBP80 are interacting with PHAX and are stabilising the CBC-PHAX complex, which explains the non-existence of a CBP80-PHAX or CBP20-PHAX complex.

3.2.4 *Structural Characterisation of the CBC-NELF-E Complex*

3.2.4.1 *Structure Determination of m^{7}GTP-CBC-NELF-E*

To visualise the binding site of NELF-E on CBC, crystallisation trials of CBCΔNLS and CBCΔNLSΔCC with different NELF-E constructs (see Appendix B.3 Table B.3) were performed in the presence of m^{7}GTP or m^{7}GpppG. Crystals were obtained only for complexes with NELF-E$^{360\text{-}380}$ and NELF-E$^{354\text{-}380}$ in the presence of m^{7}GTP or m^{7}GpppG. The crystals showed varying symmetries but in general no or only very weak density for the NELF-E peptide. The best result was obtained with CBC-NELF-E$^{360\text{-}380}$ and m^{7}GTP giving a $P2_1$ monoclinic crystal that diffracted to 2.8 Å resolution (Table 3.5). Molecular replacement was performed using the known m^{7}GpppG-CBC complex structure (PDB:1H2T, [139]) as search model. There are four m^{7}GTP bound CBC complexes in the asymmetric unit with clear electron density for the NELF-E$^{360\text{-}380}$ peptide in one of them and weaker in the other copies. The omit difference density map is shown in Appendix B.3 Figure B.4 and an overview about the data collection and refinement statistics is given in Table 3.5.

3.2.4.2 *Structure Analysis of m^{7}GTP-CBC-NELF-E*

Within m^{7}GTP-CBC-NELF-E CBC shows the same overall structure as reported for CBC bound to the cap analogue [139]. Surprisingly, The NELF-E peptide binds exactly to the same site at the CBP20 and CBP80 interface of m^{7}GTP-CBC (Figure 3.12) and in an extremely similar manner to that of ARS2 (Figure 3.9). Also the slight displacement of interacting regions of CBP20 is identical to the one described for ARS2 (see Figure 3.12). With 2501 Å^2 (PISA server [108]) the total buried surface area upon NELF-E$^{360\text{-}380}$ binding to CBC is slightly higher than for ARS2 (2257 Å^2). Furthermore, as ARS2 the NELF-E peptide shows two binding sites including residues 360-372 (proximal site) and residues 378-380 (distal site). The residues 375-377 connecting these two regions show no electron density (Appendix B.3 Figure B.4A) and thus are poorly ordered within the CBC-ARS2 complex.

There are analogous interactions of NELF-E Arg362 with CBP80 Tyr461 and CBP20 Gln55 and Asp107 (Figure 3.12B left) as previously seen for Arg854 of ARS2. Additionally, NELF-E Tyr367, that is partially stacked on Ile365, is buried in the same pocket as ARS2 Tyr859 and also the hydroxyl group of Tyr367 is interacting with the side chain of CBP20 Glu53 (Figure 3.12B middle). Within this binding site there is an

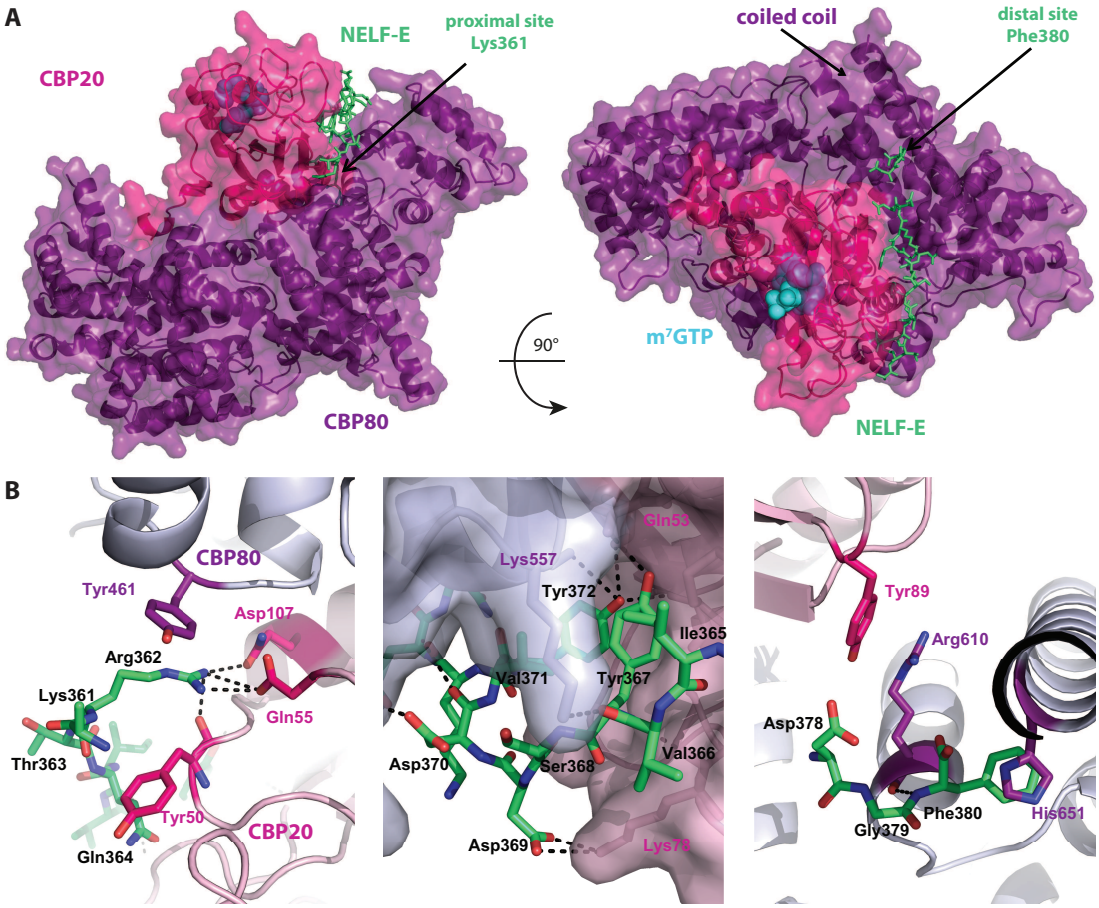


Figure 3.12 – Crystal structure of m^{7}GTP-CBCΔNLS-NELF-E$^{360\text{-}380}$

A: Overview of the m^{7}GTP-CBCΔNLS-NELF-E$^{360\text{-}380}$ crystal structure. The C-terminus of NELF-E (green) lies in the interface of CBP80 (purple) and CBP20 (pink), which binds m^{7}GTP (cyan).
B: Close-up view of the key interacting residues. Black dashed lines indicate polar contacts.

additional Tyr372, which is buried in the interface, stacking on the other side with CBP20 Ile70 and hydrogen bonding to the main-chain amino group of CBP20 Met71 (Figure 3.12B middle). Additional putative electrostatic contacts are made between Asp369 and CBP20 Lys78 and Asp370 and CBP80 Lys568. As Phe871 of ARS2 the C-terminal Phe380 of NELF-E is buried in the hydrophobic pocket of CBP80, stacking on CBP80 His651 (Figure 3.12B right). Superposition of m^{7}GTP-CBC-NELF-E with m^{7}GpppG-CBC (PDB:1H2T RMSD: 0.33 Å and 1H2U RMSD: 0.36 Å [139]) and apo-CBC (PDB:1H2V [139] RMSD: 0.58 Å) via CBP80 shows again that some residues of CBP20 are slightly shifted in the direction of the peptide, with a maximum displacement of 1.6 Å compared to 2.0 Å for the ARS2 peptide (see Appendix B.3 Figure B.4B and Figure 3.9D). Overall the interaction of NELF-E with m^{7}GTP-CBC appears slightly more robust than that of ARS2, consistent with the twenty-fold difference in affinity. An overview of all interactions between CBC and NELF-E can be found in Appendix B.3 Figure B.5.

Table 3.5 – Data collection and refinement statistics for m^{7}GTP-CBCΔNLS-NELF-E$^{360\text{-}380}$

CRYSTAL	m^{7}GTP-CBCΔNLS-NELF-E$^{360\text{-}380}$
Diffraction data	
Beamline	ID30A1
Wavelength	0.966
Space group	$P2_1$
Cell dimensions (Å)	a=113.8 b=147.2 c=153.9
Cell angles (°)	α=γ=90 β=91.48
Resolution range of data (last shell) (Å)	50.0-2.80 (2.87-2.80)
Completeness (last shell) (%)	98.4 (91.6)
R-sym (last shell) (%)	10.8 (129.2)
I/σI (last shell)	9.2 (0.91)
CC(1/2) (last shell)	0.995 (0.363)
Redundancy (last shell)	2.71 (2.70)
Refinement	
Reflections used in refinement work (free)	117089 (6345)
R-work (last shell)	0.202 (0.389)
R-free (last shell)	0.229 (0.394)
Number of non-hydrogen atoms	29303
Protein	28763 (4x CBC)
Ligand (m^{7}GTP)	132 (4x m^{7}GTP)
Peptide	408 (chains E, K, Z)
Geometry and B-factors	
RMSD (bonds)	0.009
RMSD (angles)	1.312
Ramachandran favored (%)	97.0
Ramachandran outliers (%)	0.2
Clash score	1.0103
MolProbity score	1.17
Average B-factor	74.4
Protein	74.2 (4x CBC)
Ligand (m^{7}GTP)	70.7 (4x m^{7}GTP)
Peptide	111.5 (chains E, K, Z)

3.2.5 *Comparision of ARS2 and NELE-E Binding to CBC*

Comparing the structure of CBC-ARS2 with CBC-NELF-E (see Sec. 3.2.1.2 and Sec. 3.2.4.2) and aligning the C-terminal residues of both proteins, highlights a remarkable sequence homology that was not recognised before this structural analysis (Figure 3.13A). Seven residues out of 21 are identical. Among them the main residues of both peptides involved in the interaction with CBC, namely R854 and Y859 for ARS2 and R362 and Y367 for NELF-E. Furthermore, this two residue motif (RxxxxY) is absolutely conserved in both proteins. This shows that NELF-E and ARS2 bind with their C-terminus in the same CBC binding pocket and are in fact competing for CBC binding.

In line with the high similarity of the CBC-ARS2 and CBC-NELF-E binding CBC-mut, comprising CBP80 Y461A R610E H651A and CBP20 Y50A, designed to abolish the interaction with ARS2 (see Sec. 3.2.2) showed no interaction with NELF-E$^{244\text{-}380}$ in the presence of m^{7}GTP (Figure 3.13B).

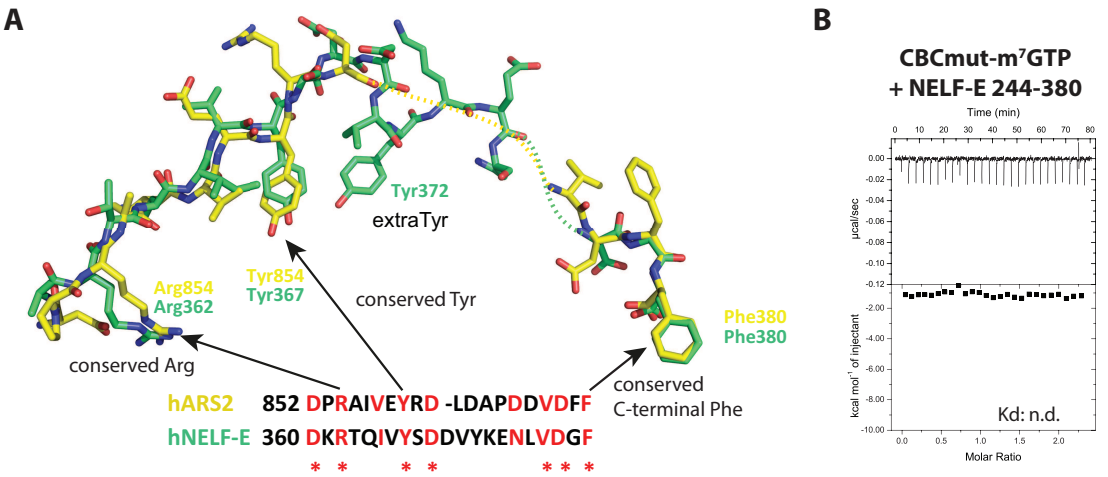


Figure 3.13 – Comparison of ARS2 and NELF-E binding on CBC

A: Comparison of the conformation of NELF-E (green) and ARS2 (yellow) C-terminal peptides after superposition of CBP80 of both CBC complexes (RMSD: 0.24 Å). Alignment of the C-terminus of NELF-E and ARS2 highlighting the conserved residues in red and the identical residues with a red star.
B: Validation of the binding site of NELF-E on CBC. In an ITC experiment the CBCmut (see Sec. 3.2.2) in the sample cell was titrated by NELF-E244-380 in the presence of m^{7}GTP. No binding was observed, whereas the wild type bound with an affinity of 50 nM (see Appendix B.3 Table B.3). The data were plotted as described in the caption to Figure 3.3C.

3.3 STRUCTURAL STUDIES OF HUMAN ARS2

3.3.1 *Structure Determination of ARS2*

To better understand the structural basics of ARS2s RNA and protein binding as well as its overall structural organisation, crystallisation trials of different ARS2 constructs alone or in the presence of RNA or the FARB peptide were performed. First crystals were obtained of ARS2$^{147\text{-}763}$ diffracting to around 20 Å. Limited proteolysis of ARS2$^{147\text{-}763}$ with trypsin followed by SEC led to crystals diffracting up to 7 Å (Figure 3.14B, C). Mass spectrometric analysis of the two double bands revealed that the protein has two separate structured regions (residues 147-274 and 393-763), which form a stable complex after cleavage. Between the double bands no difference could be detected, suggesting that the trypsin cleavage sites are very close to each other and the differences could not be captured within the peptide mass fingerprinting. The observation of two individual structured regions separated by a low complexity sequence rich in glutamates and lysines goes in line with protein disorder predictions (see Figure 3.1).

For further structural studies of ARS2 constructs containing residues 147-270 and 408-763, which lack additional glutamates compared to 147-274 and 393-763, were used. Soluble protein of these regions could only be produced when co-expressing both parts. Ultimately, crystals of ARS2$^{147\text{-}270+408\text{-}763}$ diffracting to 3.7 Å were obtained.

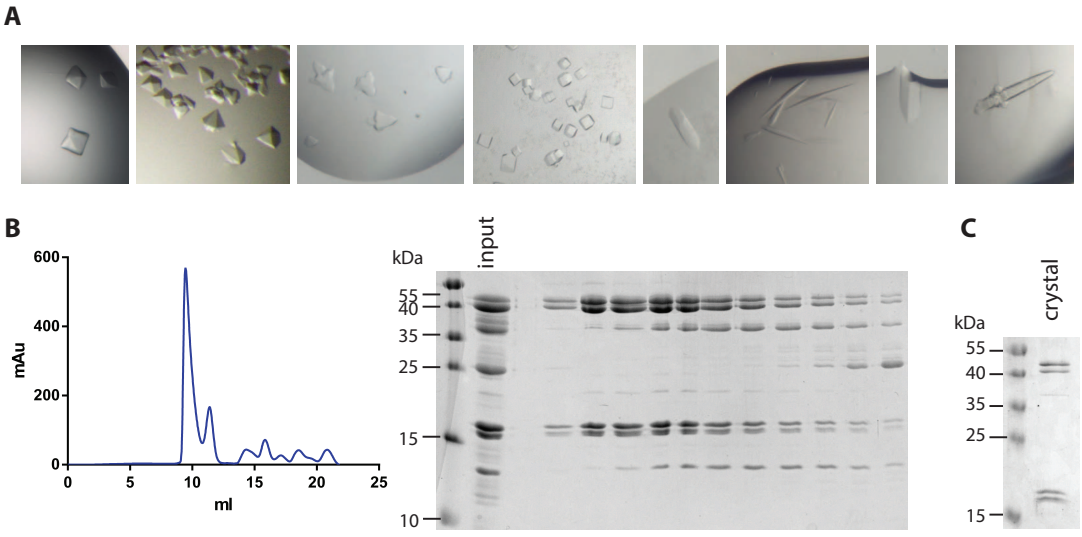


Figure 3.14 – Identification of a stable core of ARS2

A: Examples of obtained ARS2 crystals ranging from 20-200 µm.
B: Limited proteolysis revealed a stable ARS2 subcomplex. After proteolysis of ARS2$^{147\text{-}763}$ with trypsin the sample was subjected to SEC. Elution profile as well as analysis of the individual fractions by Coomassie-stained SDS-PAGE are displayed.
C: Silver-stained SDS-PAGE of a crystal obtained from crystallisation of the trypsinised complex (**B**).

Notably, first crystals diffracting up to 4 Å were obtained in the presence of the FARB peptide. Repetition and optimisation of the conditions also led to crystals diffracting to the same resolution without FARB. Using seleno-methionine-derivatised protein the structure could be solved, revealing two molecules in the asymmetric unit. Features resembling SERRATE could be visualised but the resolution was too low to permit model building. Two new crystal forms, one of which diffracted to 3.2 Å, were obtained by an additional N-terminal truncation (ARS2$^{171\text{-}270+408\text{-}763}$). The structure was solved by seleno-methionine phasing and showed one molecule in the asymmetric unit. The resultant experimental map allowed to build an almost complete model of the truncated ARS2$^{171\text{-}270+408\text{-}763}$ and ultimately of the non-truncated construct ARS2$^{147\text{-}270+408\text{-}763}$ using the 3.7 Å resolution crystal data. However, also in the presence of FARB for non of the crystals extra peptide density could be observed. Based on the obtained ARS2 structure additional constructs for crystallisation were designed lacking the unstructured residues 538-554 and residues 567-599. Using ARS2$^{171\text{-}270+408\text{-}763\Delta 567\text{-}599}$ a structure including residues 538-554 could be solved. A composite model including both loop 538-554 (from ARS2$^{147\text{-}270+408\text{-}763\Delta 567\text{-}599}$) and loop 567-599 (from ARS2$^{147\text{-}270+408\text{-}763}$) was artificially constructed and used for display purposes. All tested constructs leading to reasonable diffraction are listed in Table 3.6 and examples of crystals obtained for ARS2 are shown in Figure 3.14A. A more detailed table containing all data collection and refinement statistics can be found in the Appendix B.5 Table B.4, Table B.5 and Table B.6.

Table 3.6 – Overview of collected data from ARS2 constructs

CONSTRUCT	SPACE GROUP	CELL DIMENSIONS	RESOLUTION
ARS2 147-763	-	-	20 Å
147-763 trypsinated	$P6_4$	a=b=117.7 Å, c=147.9 Å α=β=90° γ=120°	7.0 Å
147-270 + 408-763 (+ FARB)	$P3_121$	a=b=136.64 Å, c=158.79 Å α=β=90° γ=120°	3.7 Å
171-270 + 408-763 (+ FARB)	$P6_522$	a=b=90.29 Å, c=265.72 Å α=β=90° γ=120°	3.2 Å
171-270 + 408-763 (+ FARB)	$P6_5$	a=b=105.21 Å, c=267.04 Å α=β=90° γ=120°	3.5 Å
147-270 + 408-763 Δ539-555	$P3_121$	a=b=139.58 Å, c=156.16 Å α=β=90° γ=120°	3.5 Å
147-270 + 408-763 Δ567-599	$C222_1$	a=85.51 Å, b=148.27 Å, c=235.66 Å, α=β=γ=90°	3.4 Å

3.3.2 *Structure Analysis of ARS2*

The human ARS2 structure adopts an unusual protein structure (Figure 3.15A). Only the homologous protein SERRATE from *A. thaliana* [129] shows overall the same arrangement as human ARS2 (Figure 3.15C). The core of the molecule is made up of five roughly parallel α-helices. The two extending domains, the C- and N-terminal arm, are formed by two orthogonally oriented α-helices containing the N- and C-

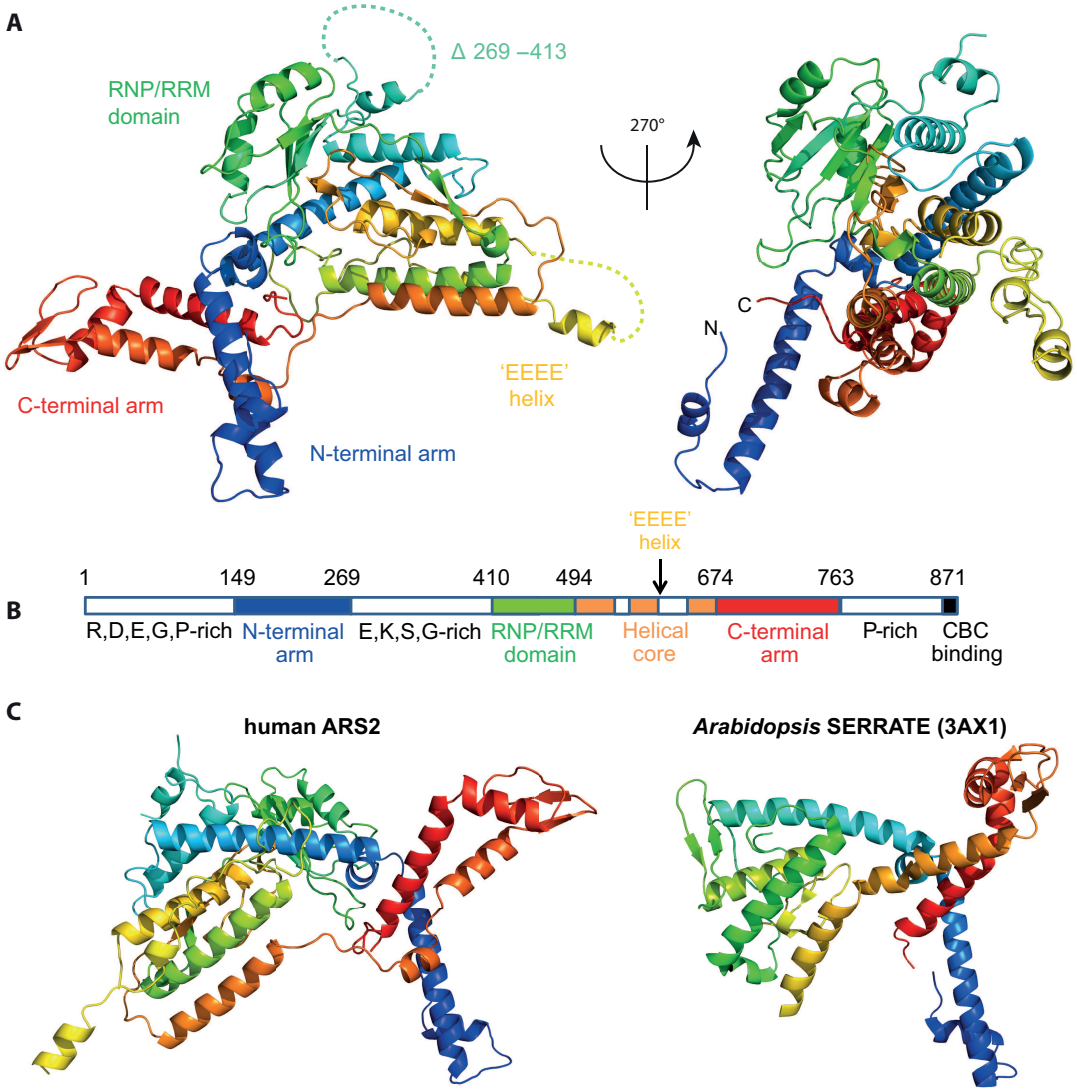


Figure 3.15 – Crystal structure of human ARS2

A: Composite ARS2 model showing loop 538-554 and 'EEEE' helix from loop 567-599. The structure is rainbow coloured from blue (N-terminus) to red (C-terminus). Loop regions are indicated by dashed lines.
B: Schematic diagram showing the ordered regions that are separated by presumed intrinsically disordered regions.
C: Comparison of human ARS2 and SERRATE crystal structure. Structures are coloured as described in (**A**).

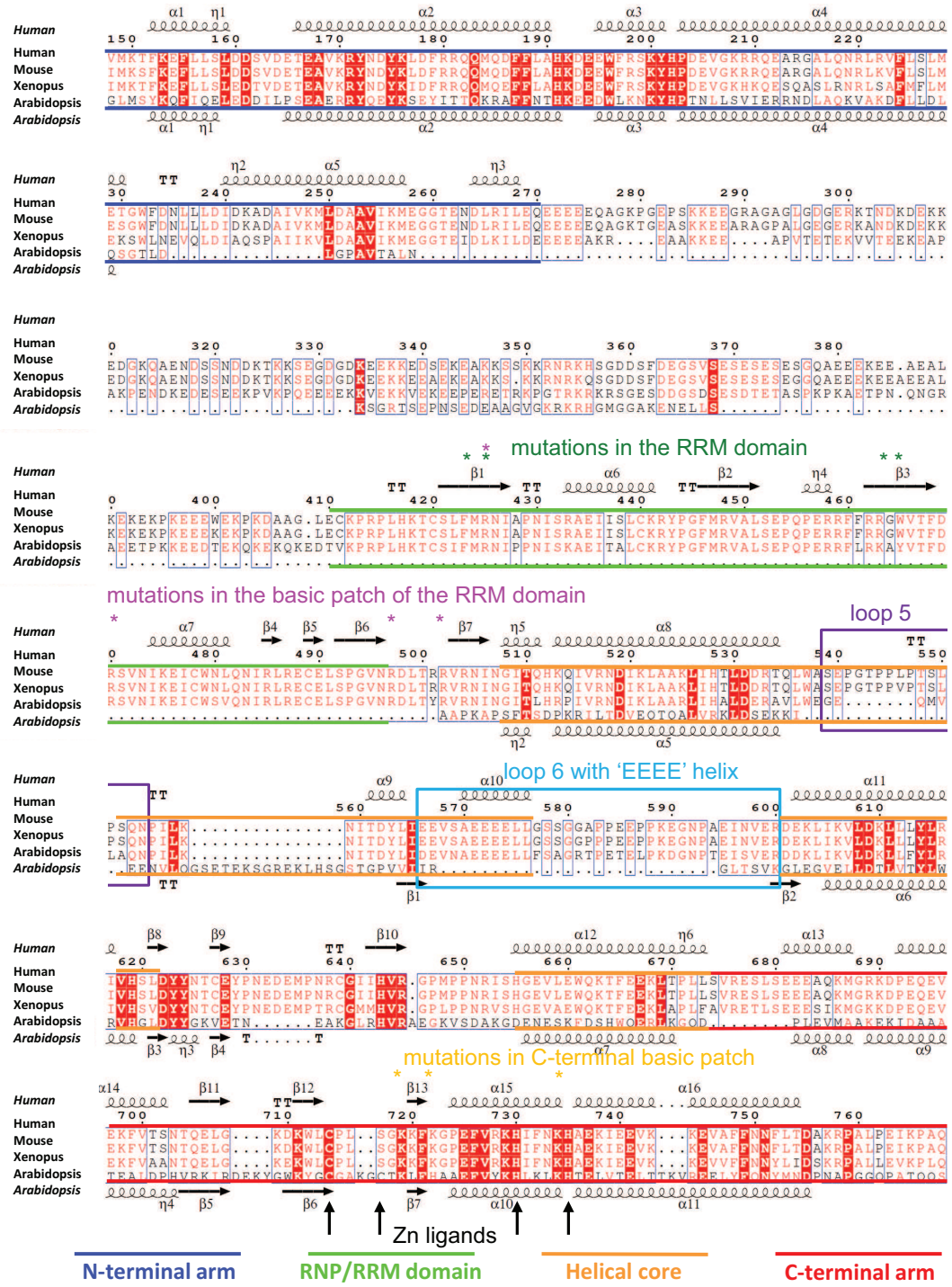


Figure 3.16 – Multiple alignment of ARS2 protein sequences and structures

Bioinformatical alignment of the protein structure of human ARS2 and SERRATE (PDB: 3AX1 [129]) as well as their protein sequences and the ARS2 protein sequence from mouse and xenopus. The human ARS2 domain organisation, the loop 538-554, the loop 567-599 containing the 'EEEE' helix, residues involves in zinc binding in *A. thaliana* and mutated lysines in this domain as well as mutations introduced in the RRM domain of hARS2 are highlighted.

terminus of the crystallised construct. The metazoan specific RRM or RNP domain is packed on the helical core and shows the typical βαββαβ fold. The presence of disordered loops 538-554 and 567-599 as well as the disordered N- and C-terminus and the residues 270-408 implicate their possible roles in interactions with other proteins and/or RNA. The main difference to the SERRATE structure is that the core is reinforced by additional helices and the inserted metazoan specific RRM domain (Figure 3.15C). Furthermore, the C-terminal arm of human ARS2 is rotated around 90° compared to SERRATE and is not stabilised by a bound zinc ion, although the folding is similar. An alignment of the sequences as well as the secondary structures of human ARS2 and SERRATE is displayed in Figure 3.16.

The RRM domain sits as distinct domain on top of the helical core, making only few hydrophobic and polar contacts to the rest of the molecule. Four regions of the core of ARS2 (187-200, 251-269, 631-641 and 759-762) are holding the RRM domain in place. The helical surface is solvent exposed, while the β-sheets are oriented towards the helical core, forming a grove between these two domains (Figure 3.15A). Interestingly, this region shows a positively charged surface (Figure 3.18A), consisting mainly of the β-sheets of the RRM domain. Residue Phe423, Arg425, Arg463 and Trp465 from the second and third β-sheet are pointing towards the solvent, indicating a putative interaction surface (Figure 3.17A).

Comparing human ARS2 RRM domain with other RRM domains it was found that the RRM of RBM7 (PDB: 5LXY [53]), which binds ZCCHC8 within the nuclear exosome targeting (NEXT) complex, shows the highest structural similarity. Superposition of the two structures shows that the hydrophobic patch of RMB7 comprising residues Leu29, Leu70, Ile73 and Leu75 involved in ZCCHC8 binding is also present in ARS2 (Leu439, Leu481, Ile484 and Leu486) (Figure 3.17B). Thus, the RRM domain of ARS2 could indeed be involved in RNA and/or protein binding.

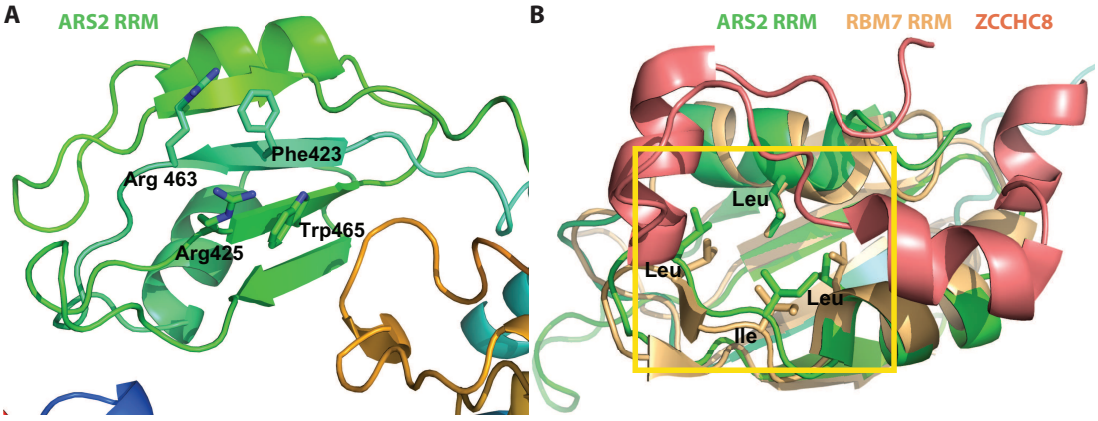


Figure 3.17 – Structure of ARS2's RRM domain

A: Zoom-in view of the RRM domain showing residues putative involved in protein or RNA interaction as sticks.
B: Superposition of ARS2 RRM domain with the RRM domain of RBM7 bound to a ZCCHC8 peptide (PDB: 5LXY RMSD: 1.38 Å [53]). Identical residues forming the hydrophobic pocket (yellow box) are shown as sticks.

However, within most crystal forms the basic RRM domain surface is the binding site of a negatively charged helix ('EEEE' helix) formed by residues 572-EEEE-575 from a neighbouring molecule (Figure 3.18B). This 'EEEE' helix occurs in the highly conserved and metazoan specific solvent exposed loop 567-599 (Figure 3.16, Figure 3.18). Remarkably, irrespective of the space group and crystal form, this intermolecular crystal contact is present. Thus, to accommodate the same crystal contact within different packing environments, some angular flexibility of the loop is required.

Additionally to the loop 567-599, which shows with the 'EEEE' helix a structural component and could partially be built, the loop 539-554 was in most crystal forms disordered probably due to flexibility, except within a crystal form of

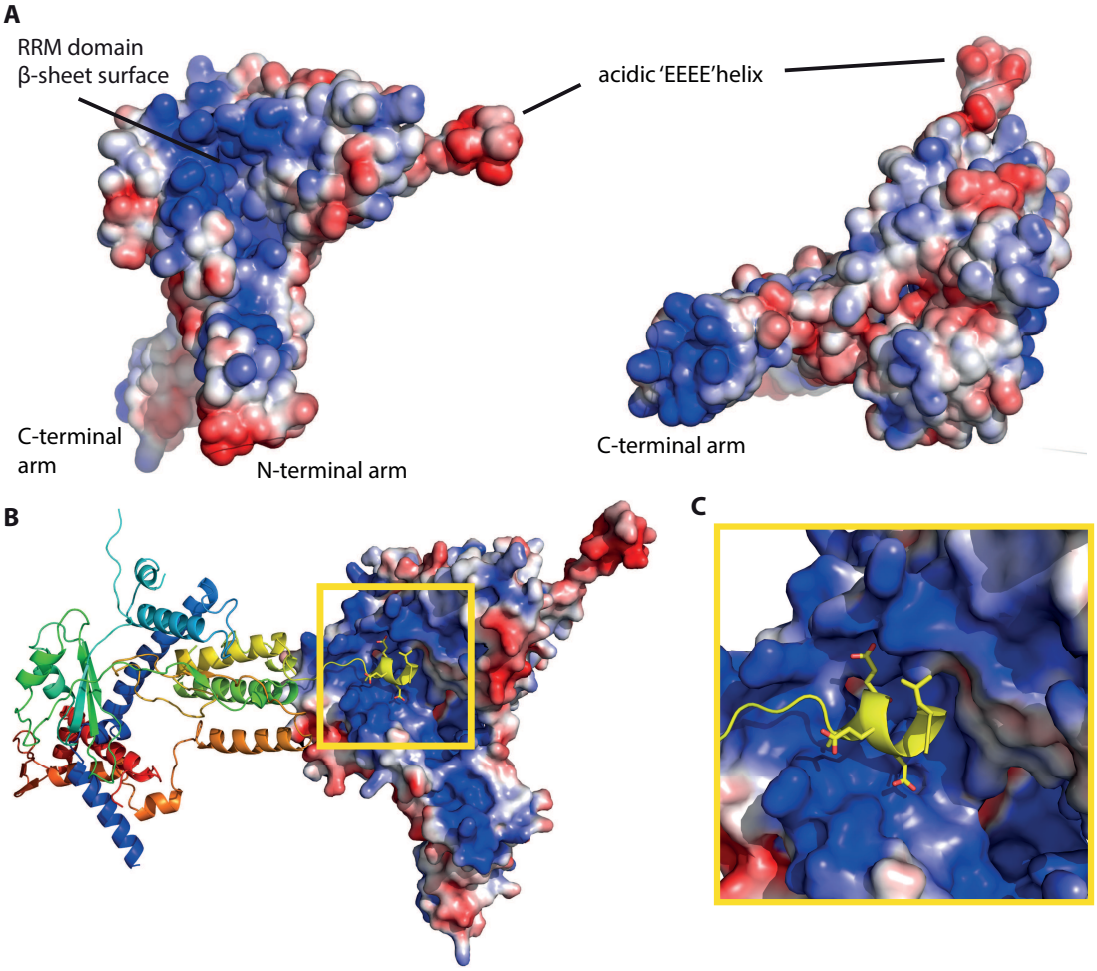


Figure 3.18 – Electrostatic surface potential of ARS2 and ARS2 crystal packing

A: Electrostatic surface potential (red –ve, blue +ve) for human ARS2$^{147\text{-}270+408\text{-}763}$ in two orientations.
B: Crystal contact between the 'EEEE' helix and the basic RRM β-sheet surface of a neighbouring molecule within an ARS2$^{171\text{-}270+408\text{-}763}$ crystal.
C: Close-up view of the crystal contact made by acidic 'EEEE' helix binding onto the basic RRM β-sheet surface (**B**).

ARS2$^{147\text{-}270+408\text{-}763\Delta 567\text{-}599}$, in which the loop is well-structured, likely due to crystal packing constraints. Notably, this loop contains Thr543, which is reported to be a major phosphorylation site of human ARS2 [85].

Another positively charged surface, apart from the RRM domain, is located at the outer surface of the C-terminal arm (Figure 3.18A), which in *A. thaliana* contains a zinc finger. Although the C-terminal arm shows a rotation of around 90° its overall structure apart from the missing zinc ion is very similar. Three out of the four residues involved in zinc binding in SERRATE are present in ARS2 (Cys714, His730 and His735), whereas the Cys505 in SERRATE is replaced by Ser717 (Figure 3.19). Furthermore, the C-terminal arm of ARS2 contains many lysines, especially at the tip of the arm (Lys709, Lys711, Lys719, Lys720, Lys722 and Lys734) (Figure 3.18).

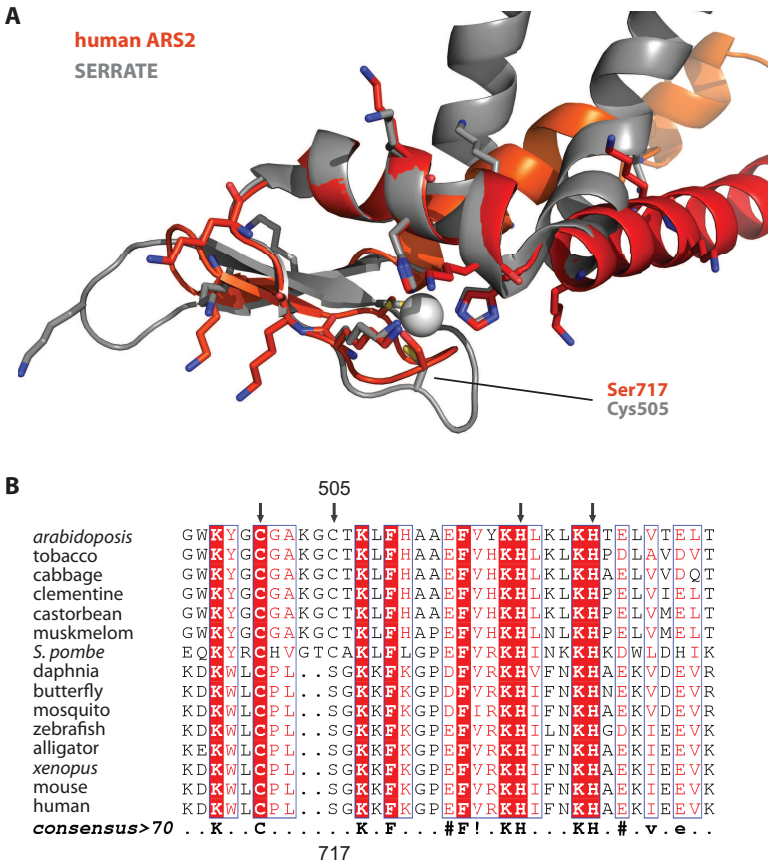


Figure 3.19 – Comparison of the C-terminal domain of ARS2 and SERRATE

A: Overlay of the C-terminal arm of ARS2 (orange) on the zinc finger of *A. thaliana* SERRATE (gray) (PDB: 3AX1 [129]) showing the residues involved in zinc binding for SERRATE as well as all lysines as sticks.
B: Multiple sequence alignment of amino acids of the N-terminal arm of ARS2. Conserved residues are highlighted in red, the residues involved in zinc binding within SERRATE are indicated with arrows and the residue Cys505 of SERRATE and Ser717 of human ARS2 are labeled.

3.4 INTERACTION AND FUNCTIONAL STUDIES OF ARS2

Initially, different ARS2 constructs designed based on protein disorder prediction and limited proteolysis were purified and tested for their affinity towards CBC (see Sec. 3.1.2), RNA and FLASH. After obtaining the crystal structure of ARS2 (see Sec. 3.3.2) additional specific mutants were constructed and their binding properties analysed. An overview of the different constructs and mutants used to identify the interacting regions of ARS2 with RNA and FLASH is given in Table 3.7.

Table 3.7 – ARS2 constructs

CONSTRUCT	COMMENT
147-871	N-terminal truncation of predicted disordered residues
147-763	N- and C-terminal truncation of predicted disordered residues
147-270 + 408-763	predicted ordered residues forming a stable complex
147-270 + 494-763	as above, plus additional deletion of the RRM domain
147-871 K719A K722A K734A	mutations in the C-terminal basic patch
147-871 Δ692-743 GSG	deletion of the C-terminal arm, replaced by GSG
147-871 Δ271-408 GSGSGSGS	N-terminal truncation and deletion of predicted disordered residues in the mid domain, replaced by GSGSGSGS
147-871 R426G R470G R498A R502A	mutations within the RRM domain
147-871 R424A R426G R463A R465A	mutations of the basic patch within the RRM domain
147-871 Δ539-554 GSA	deletion of loop 539-554, replaced by GSA
147-871 Δ539-554 AAA	deletion of loop 539-554, replaced by AAA
147-871 Δ539-554 15A	mutation of loop 539-554 to alanine
147-871 Δ567-599 GSGSGS	deletion of loop 567-599, replaced by GSGSGS

3.4.1 *ARS2 RNA Binding*

To study the RNA binding of ARS2 fluorescence polarisation experiments were performed using FAM-labeled RNA. ARS2$^{147\text{-}871}$ showed affinity for single stranded (ss) RNA (K_D~2-5 µM) and partially double stranded (pds) RNA (K_D~3-5 µM), whereas only a very weak affinity (K_D~25-37 µM) for double stranded (ds) RNA could be determined (Figure 3.20). Although different RNA sequences were used within the fluorescence polarisation assay, no sequence specificity could be observed (data not shown). ARS2$^{147\text{-}763}$ showed the same binding behaviour, whereas ARS2$^{147\text{-}270+408\text{-}763}$ seemed to bind some pdsRNA with much lower affinity (K_D~3-5 µM or 30-50 µM depending on the pdsRNA, data not shown) than ssRNA (K_D~2-5 µM). However, so far no obvious explanation for the differences could be made and further studies need to be performed to better understand ARS2 RNA binding.

To determine the RNA binding site on ARS2 different constructs as well as ARS2$^{147\text{-}871}$ point mutants based on the crystal structure were used and their affinities for ssRNA were analysed. For ARS2 lacking the RRM domain (ARS2$^{147\text{-}270+494\text{-}763}$) only a very weak RNA binding was observed (K_D~46 µM) (Figure 3.20B). Also point mutations in the basic patch in the RRM domain (ARS2 F424A R426A R463A Y465A) and in the basic batch just next to it (R426G R470G R498A R502A) led to a ten times weaker RNA affinity as the wild type (Figure 3.20B), indicating that this region is important for RNA binding by ARS2.

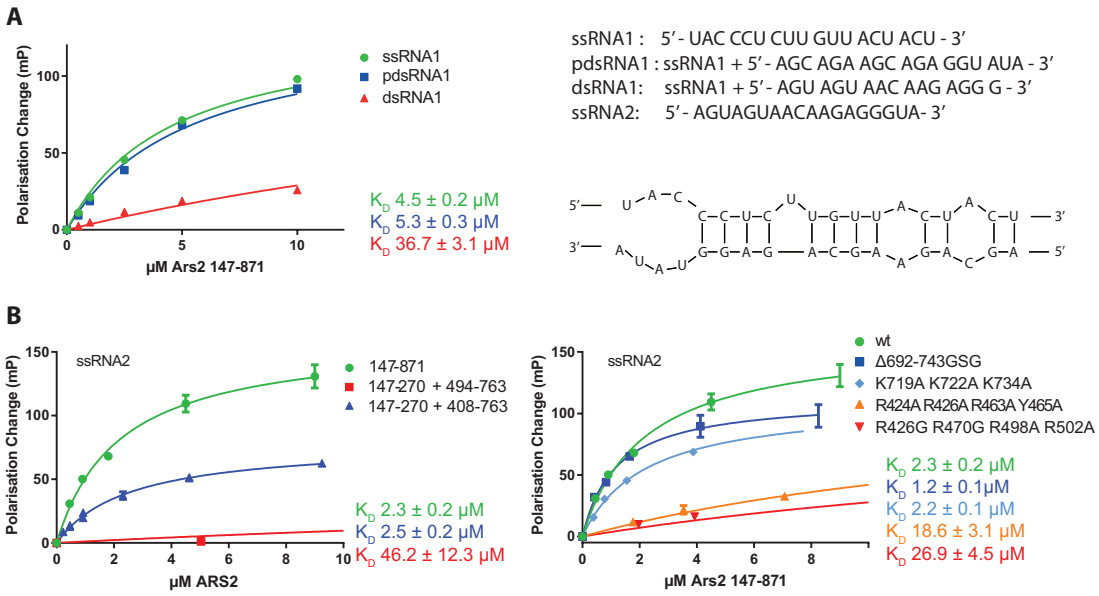


Figure 3.20 – ARS2 RNA binding

A: Single stranded (ss), partially double stranded (pds) and double stranded (ds) RNA were tested for ARS2 binding via fluorescence polarisation. FAM-labeled RNA was titrated with increasing amounts of ARS2 and the change in fluorescence polarisation measured. Mean values and standard deviations represent the average from three independent experiments.
B: Different ARS2 constructs and mutants were used within the fluorescence polarisation assay to map the RNA binding site. Set up as described in A using only ssRNA.

3.4.2 *ARS2 – FLASH Interaction*

In the transcription of histone pre-mRNAs FLICE-associated huge protein (FLASH, 1962 residues) plays an important scaffolding role for processing factors. It was shown that a central 13 amino acid peptide of FLASH (931-943 DELEEGEIRSDSE), called FLASH-ARS2-binding (FARB) region, is sufficient for the interaction with ARS2 [104]. Furthermore, this interaction has been reported to be important for S phase progression [104]. To better understand the basis of the ARS2-FLASH interaction the structural, biochemical and functional studies were carried out.

3.4.2.1 *ARS2 – FLASH Structural Characterisation*

Crystallisation trials of different ARS2 constructs with the 13 amino acids long FARB peptide or with a 26 amino acids long peptide including the FARB sequence ($FLASH^{921\text{-}946}$) yielded various crystals in many different conditions. The structures were solved by molecular replacement using the obtained ARS2 structure as search model. However, only for crystals of $ARS2^{147\text{-}270+408\text{-}763}$ plus $FLASH^{921\text{-}946}$ that diffracted up to 6 Å additional density possibly belonging to the peptide could be found. This additional density lies between the end of the C-terminal arm of $ARS2^{719\text{-}734}$ and $ARS2^{408}$ (data not shown) and could be a part of the FARB peptide. As FARB is a very acidic peptide it is possible that it interacts with the basic patch of the C-terminal arm of ARS2 (see Figure 3.18A).

3.4.2.2 *ARS2 – FLASH Biochemical Characterisation*

By ITC the minimal ARS2 construct $ARS2^{147\text{-}270+494\text{-}763}$, missing the RRM, bound to FARB with almost the same affinity (K_D~0.5 µM) as $ARS2^{147\text{-}270+408\text{-}763}$ (K_D~0.6 µM) (Figure 3.21, Table 3.8). ITC experiments with $ARS2^{147\text{-}871}$ were performed at higher salt conditions (250 mM NaCl), which enabled the use of $ARS2^{147\text{-}871}$ and resulted in a K_D of around 5 µM for the FARB and $FLASH^{921\text{-}946}$ peptide (Table 3.8). To map the FLASH binding site of ARS2 different ARS2 constructs were designed based on the solved crystal structure (see Table 3.7). In order to perform pull down experiments between FARB and ARS2, $FLASH^{903\text{-}943}$ containing FARB ($FLASH^{931\text{-}943}$ – SDLNKENQKPIYKSDKCTEADTCKNSPL*DELEEGEIRSDSE*) was cloned into the GST-vector pETM30 and expressed in *E. coli*. After immobilizing GST-$FLASH^{903\text{-}943}$ on the Glutathione Sepharose and incubating with the different $ARS2^{147\text{-}871}$ constructs the elutions were analysed, showing that $ARS2^{\Delta 692\text{-}743\ GSG}$ and $ARS2^{K719A\ K722A\ K734A}$ did not interact with GST-$FLASH^{903\text{-}943}$ (Figure 3.21B). ITC experiments of these two constructs confirmed that both mutations abolished the interaction with the FARB peptide (Table 3.8). Thus, the deletion of ARS2's C-terminal arm as well as the mutagenesis of its basic patch resulted in the loss of FARB binding. To exclude unspecific binding of the acidic FARB peptide to the basic patch of ARS2 a scrambled ARS2 peptide (EGEIESELSEDDR), containing the same residues randomly aligned and therefore showing the same acidity as the original FARB peptide, was used as control resulting in no measurable interaction with $ARS2^{147\text{-}871}$ by ITC (Figure 3.21A).

Furthermore, ARS2$^{634\text{-}871}$, ARS2$^{634\text{-}763}$ and ARS2$^{677\text{-}758}$ were able to bind FARB with 10 µM affinity, whereas ARS2$^{634\text{-}763K719A\ K722A\ K734A}$ showed no binding (Table 3.8). Also here, the scrambled FARB peptide showed no binding to ARS2$^{677\text{-}758}$ by ITC and supports a specific interaction between the peptide and the lysines located in the C-terminal arm of ARS2. However, analyses of the short constructs alone by 1D NMR gave no evidence for the existence of structured regions within this construct (data not shown).

The result that lysines located in the C-terminal arm of ARS2 are mediating the interaction with FLASH goes in line with the observed additional density close to ARS2$^{719\text{-}734}$ (see Sec. 3.4.2.1).

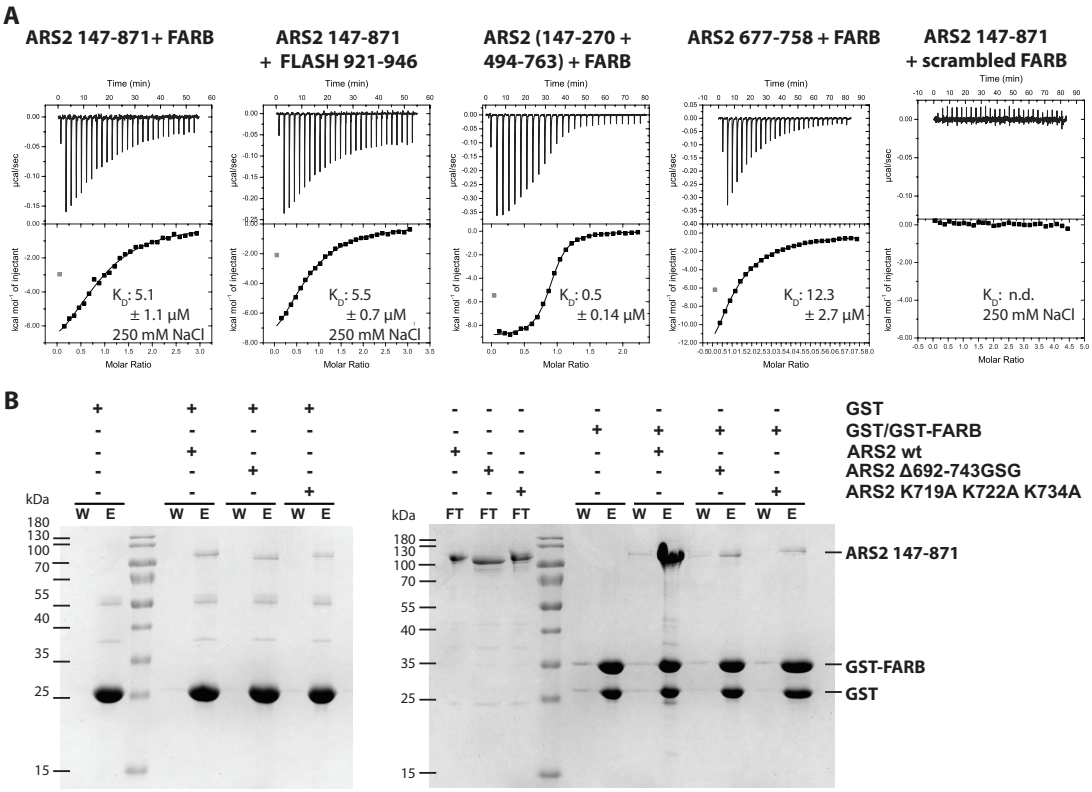


Figure 3.21 – Interaction of ARS2 with FLASH

A: Determination of K_D between ARS2 and FLASH constructs. ITC experiments with ARS2 in the sample cell and FLASH in the syringe were used to measure the affinity of different constructs. The data were plotted as described in the caption to Figure 3.3C.
B: Mapping the interaction region of FLASH on ARS2. GST or GST-FLASH$^{903\text{-}943}$, which includes the FARB peptide, was immobilised on Glutathione resin and incubated with different ARS2$^{147\text{-}871}$ constructs. The flow through (FT), last wash (W) and elution (E) fractions were analysed via Coomassie-stained SDS-PAGE.

Table 3.8 – ARS2 - FLASH interactions

Using ITC the interactions of different ARS2 constructs to FLASH were tested.

INTERACTION	K_D
ARS2 147-871 + FARB (250 mM NaCl)	5.1 ± 1.1 µM
ARS2 147-871 + FLASH 921-946 (250 mM NaCl)	5.5 ± 0.7 µM
ARS2 147-871 + scrambled FARB (250 mM NaCl)	n.d.
ARS2 147-763 + FARB (250 mM NaCl)	2.1 ± 0.4 µM
ARS2 (147-270 + 408-763) + FARB	0.6 ± 0.1 µM
ARS2 (171-270 + 408-763) + FARB (250 mM NaCl)	3.6 ± 0.3 µM
ARS2 (147-270 + 494-763) + FARB	0.5 ± 0.14 µM
ARS2 634-871 + FARB	11.1 ± 4.8 µM
ARS2 634-763 + FARB	10.8 ± 2.2 µM
ARS2 677-758 + FARB	12.3 ± 2.7 µM
ARS2 677-758 + scrambled FARB	n.d.
ARS2 147-871 K719A K722A K734A + FARB (250 mM NaCl)	n.d.
ARS2 147-871 Δ692-743 GSG + FARB (250 mM NaCl)	n.d.
ARS2 634-763 K719A K722A K734A + FARB	n.d.

3.4.3 *ARS2 Cell Based Assays*

To gain a better insight into the functions of the different ARS2 domains and loops constructs based on the crystal structure (see Table 3.7) were expressed as EGFP-tagged fusion protein in HEK 293T/17 cells. Therefore, EGFP followed by a GSGGGS linker and the ARS2 gene were cloned into pcDNA3.1. The EGFP tag enabled to visualise directly the transfection efficiency and localisation of the proteins. All constructs except the N-terminal truncation ARS2$^{147\text{-}871}$ and the deletion of loop 538-554 localised in the nucleus as the wild type (Figure 3.22). ARS2$^{147\text{-}871}$ showed only cytoplasmic localisation, which is in agreement with the described NLS within these N-terminal residues [114, 160]. The deletions of loop 538-554 (EGFP-ARS2$^{\Delta 538\text{-}554\ GSA}$ and EGFP-ARS2$^{\Delta 538\text{-}554\ AAA}$) but not the mutagenesis of all residues to alanine (EGFP-ARS2$^{\Delta 538\text{-}554\ 15A}$) led to a localisation of EGFP-ARS2 in distinct subnuclear compartments, which are likely to be the nucleoli. Since overexpression of mCherry-Nop58 or coilin, two predominantly nucleoli localised proteins, showed no specific localisation in the nucleus within the performed experiments (data not shown), no conclusions about the subnuclear compartment could be made. Interestingly, ARS2 wild type however seemed not to localise within this nuclear regions.

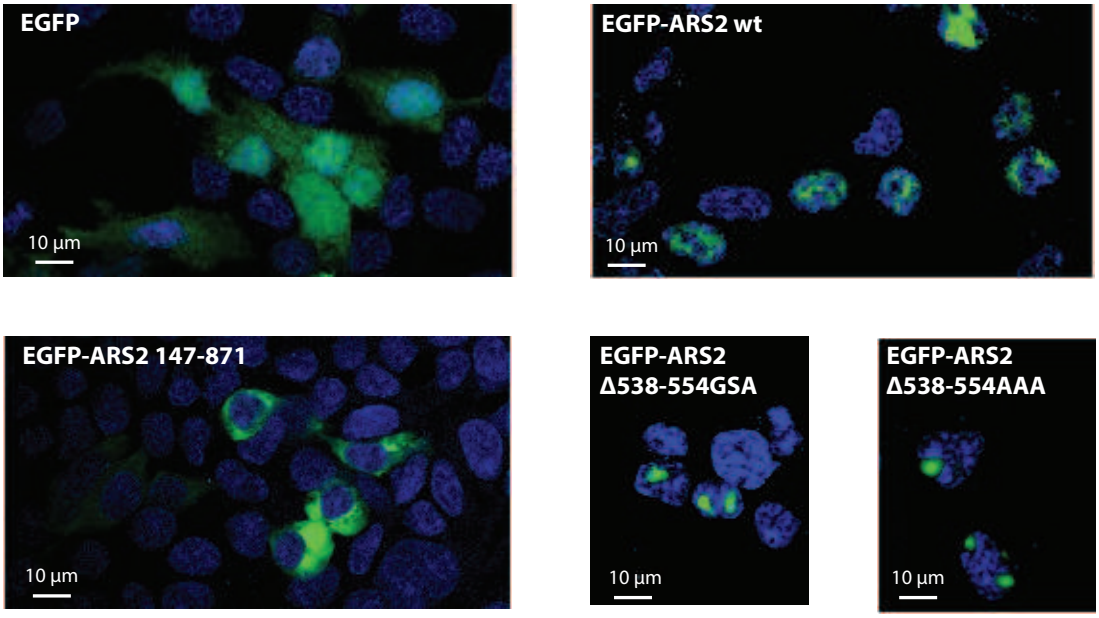


Figure 3.22 – Localisation of EGFP-ARS2 in HEK293T/17 cells

30 h post transfection cells were fixed, permeablised and stained with DAPI before imaging by confocal microscopy. Pictures taken for the DAPI (blue) and EGFP (green) channel were overlaid and the contrast increased for printing purpose.

3.4.3.1 *Cell Cycle Analysis*

Recently, it was reported by O'Sullivan *et al.* that overexpression of ARS2 leads to early S phase cell cycle arrest [160]. In the study they also predicted the structure of murine ARS2 and analysed the effect of different mutations based on the prediction. Having solved the human ARS2 structure allowed the identification of the exact residues involved in RNA (see Sec. 3.4.1) and FLASH binding (see Sec. 3.4.2). To broaden the studies the effect of the different mutants on the cell cycle was analysed. Therefore, 48 h post transfection of the EGFP-tagged ARS2 constructs the cells were pulsed with BrdU and analysed for their BrdU incorporation and DNA content in respect to their EGFP signal. Looking at the distribution of the cells among the cell cycle a slight not significant increase of cells in the S phase was detected for EGFP-ARS2 positive cells compared to cells transfected with the empty plasmid or EGFP (data not shown). The cells showed a normal cell phase distribution (Figure 3.23 left). Analysis of the different ARS2 constructs showed that the substitution of residues 538-554 by GSA or AAA led to an accumulation of cells in late S Phase or G2/M phase (Figure 3.23 right, Figure 3.24B). Consequently, fewer cells were in G0/G1 or S phase. Interestingly, the mutation of all these residues to alanine (Δ538-554 15A) did not show this behavior (Figure 3.24B-D). This is in accordance with the differences observed for the cellular localisation of these constructs (see Sec. 3.4.3) and needs further investigation to exclude possible misfolding and/or mislocalisation of the mutants with the truncated loop 538-554.

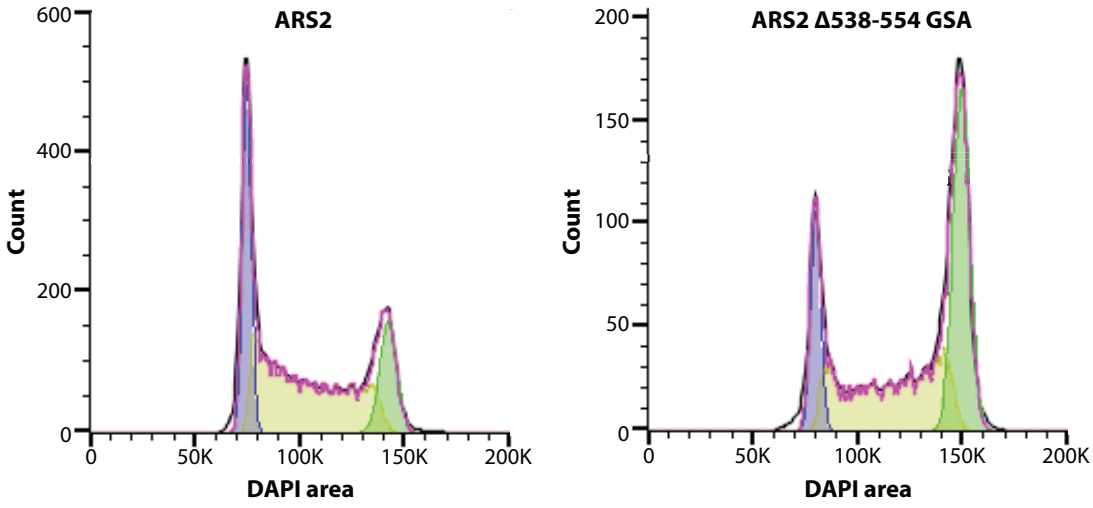


Figure 3.23 – Effect of ARS2 overexpression on the cell cycle

Cell cycle analysis after overexpression of ARS2 and ARS2$^{\Delta 538\text{-}554\ GSA}$. HEK 293T/17 cells were transfected with EGFP-ARS2 constructs and analysed after 48 h by flow cytometry for their EGFP and DNA content. Displayed is the DNA content of EGFP positive cells transfected with EGFP-ARS2 (left) and EGFP-ARS2$^{\Delta 538\text{-}554\ GSA}$ (right). The data were fitted with the univariate model to quantify the G0/G1 (blue), S (yellow) and G2/M (green) phase populations.

The mutant ARS2$^{R426A\ R470A\ R498A\ R502A}$, with mutations within the basic patch around the RRM domain that reduced RNA binding (see Sec. 3.4.1), showed a significant enrichment in cells within the S phase compared to the wild type (Figure 3.24C). As less cells were in the G0/G1 phase, this could indicate a possible involvement of these residues in S phase progression and needs additional analysis.

Although O'Sullivan *et al.* described an early S phase arrest of C2C12 cells when overexpressing ARS2, within the performed experiments using HEK 293T/17 only a very small effect of ARS2 overexpression could be observed. This might explain that only the deletion of loop Δ538-554 and mutations within the basic patch around the RRM led to significant difference from the wild type. Until now the function of loop Δ538-554 is unknown. Residues R42,6 R470, R498 and R502A were shown to mediate

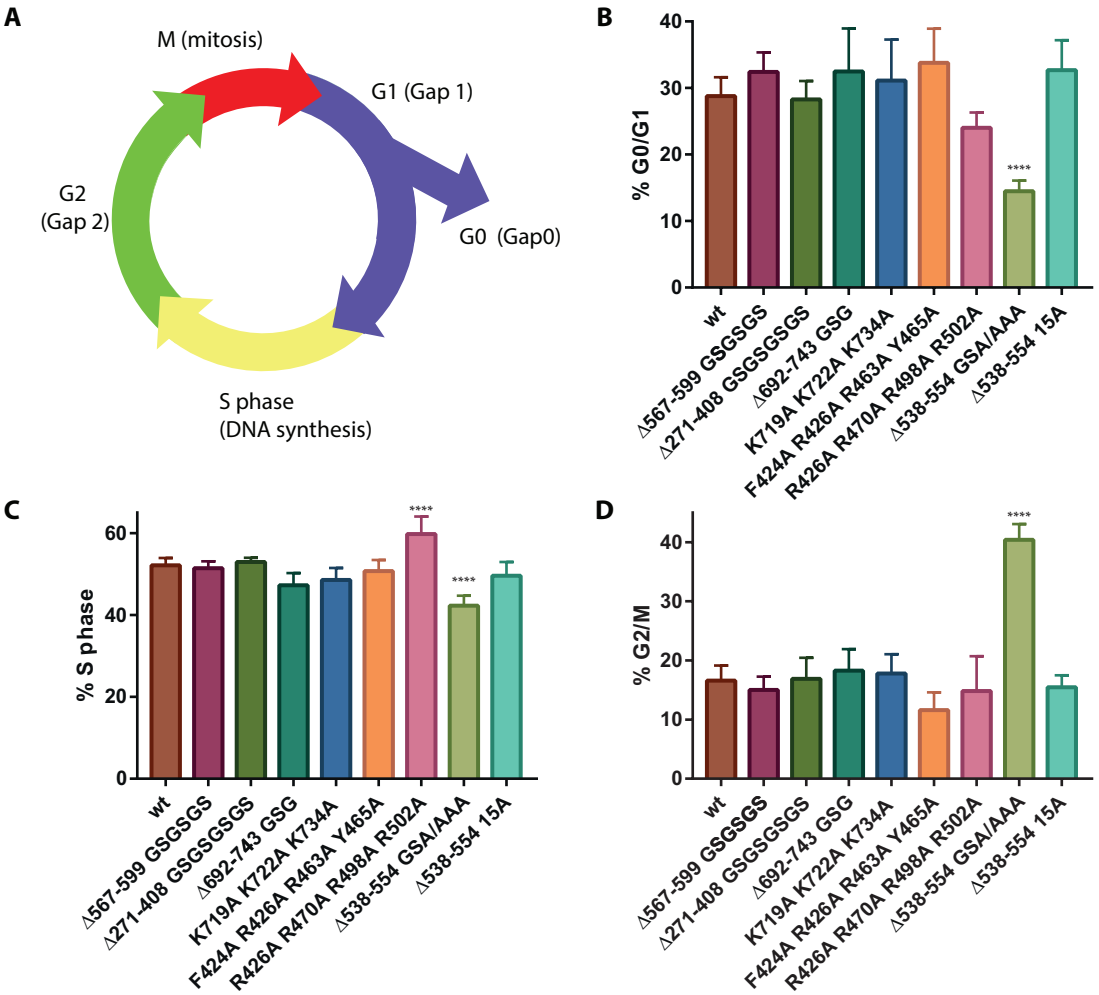


Figure 3.24 – Effect of overexpression of different ARS2 constructs on the cell cycle

A: Schematic overview of the cell cycle.
B-D: Quantification of EGFP positive cells in each phase of the cell cycle: (**B**) G0/G1 phase, (**C**) S phase and (**D**) G2/M phase. Values derived from curve fitting described in caption of Figure 3.23. P-value **** < 0.0001.

RNA binding, which could impact S phase progression. However, the mutant with mutations in the RRM domain, that also has reduced RNA affinity, did not show the same cell phase distribution. The C-terminal domain mutants (ARS2$^{\Delta 692\text{-}743\ GSG}$ and ARS2$^{K719A\ K722A\ K734A}$) involved in FLASH binding (see Sec. 3.4.2.2) showed no effect on S phase progression. Thus, no conclusion of the impact of residues involved in RNA- or FLASH-binding on the cell cycle could be made.

3.4.3.2 *Identification of ARS2 Interaction Partners*

To identify regions of ARS2 involved in protein-protein interactions EGFP-tagged ARS2 constructs and their interacting factors were isolated via affinity purification (AP) using GFP-Trap (see Appendix B Figure B.10). EGFP-ARS2$^{\Delta 538\text{-}554\ GSA}$, which localised in a subnuclear compartment (see Sec. 3.4.3), remained in the insoluble fraction and was not captured by GFP-Trap (see Appendix B Figure B.10). Therefore, this sample together with EGFP and an empty vector was used as control. Tandem mass tag (TMT) labeling and mass spectrometric analysis of the eluted fractions was performed by the EMBL Proteomics Core Facility. The results (Figure 3.25) showed an enrichment of CBP80, CBP20 and PHAX for ARS2 wild type as described by previous studies [73]. Furthermore, many proteins involved in RNA processing could be identified such as splicing factors (THRAP3, TRA2B, SNRNP70 and RBMX), helicases (DDX5, DDX3X, DHX30 and DDX17), 3' end processing factors (CPSF2 and CPSF3) and proteins involved in RNA packaging (different hnRNP complex components). Besides PHAX also hnRNP C as well as ALY and components of the TREX complex (THOC1 and THOC2), which are all associated with the transport of RNAs, were enriched for ARS2. Also different importins mediating protein import into the nucleus were identified (importin-α1, importin-α3, importin-α4, importin-α7, importin-11 and importin-β1). In high amounts NCBP3, WDR82 and the zinc finger proteins ZC3H4, ZCCHC8 and ZFC3H1 were determined as ARS2 associated proteins. Notably, also Gebhardt *et al.* reported the association of the nuclear cap binding protein NCBP3 with ARS2 [64]. WDR82 is a CTD binding protein involved in histone H3K4 methylation [115, 223]. Until now the function of ZC3H4 is unknown, whereas ZCCHC8 is part of the NEXT and ZFC3H1 belongs to the PAXT complex. Interestingly, the interaction of these two complexes is mutually exclusive and mediated via MTR4 [144], which within this study was also identified as ARS2 associated protein.

Comparing the different ARS2 constructs many differences between ARS2 wild type and ARS2$^{\Delta 692\text{-}743\ GSG}$ and ARS2$^{K719A\ K722A\ K734A}$ were observed (Figure 3.25C, D). Both constructs, ARS2$^{\Delta 692\text{-}743GSG}$ which is missing the basic C-terminal arm as well as ARS2$^{K719A\ K722A\ K734A}$ containing mutations in the basic patch of the C-terminal arm, showed for example no enrichment in NCBP3, ZC3H4, ZFC3H1, PHAX and MTR4 compared to ARS2 wild type. These results indicate that the C-terminal arm of ARS2 mediates different protein-protein interactions and is important for the assembly of correct RNA processing complexes.

All as significantly enriched classified proteins for EGFP-ARS2 compared to EGFP and the two C-terminal domain mutants, EGFP-ARS2$^{K719A\ K722A\ K734A}$ and EGFP-ARS2$^{\Delta 692\text{-}743GSG}$, are listed in Appendix B.6 Table B.7.

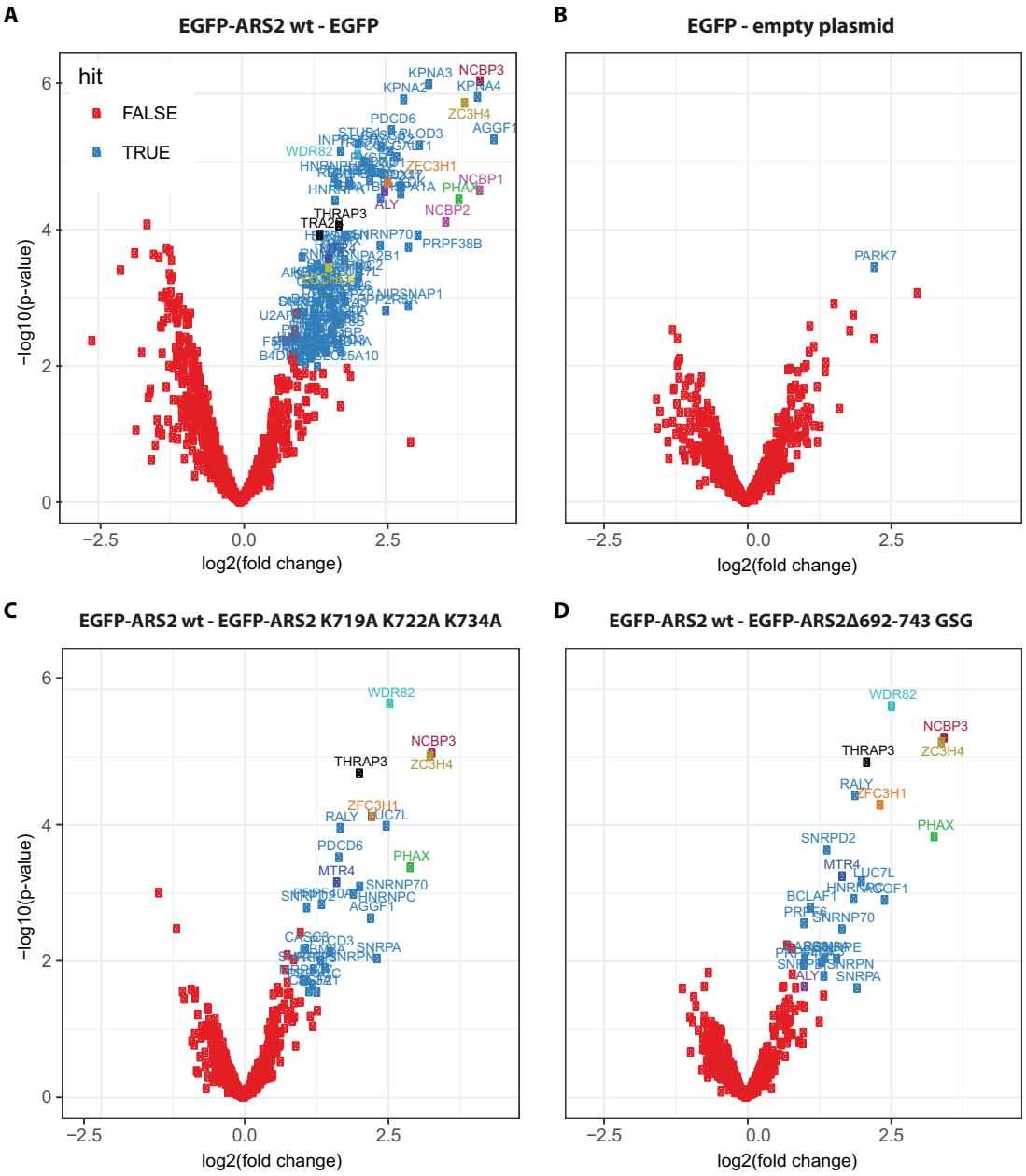


Figure 3.25 – Proteomic analysis of proteins associated with ARS2

Volcano plot comparing the enrichment of proteins identified for
A: EGFP-ARS2 wild type with EGFP.
B: EGFP with empty plasmid.
C: EGFP-ARS2 wild type with EGFP-ARS2$^{K719A\ K722A\ K734A}$.
D: EGFP-ARS2 wild type with EGFP-ARS2$^{\Delta 692\text{-}743\ GSG}$.
The as true classified hits are listed in Appendix B.6 Table B.7.

3.5 NCBP3 BINDING STUDIES

As NCBP3 was identified as binding partner of ARS2 by EGFP-ARS2 pull down (see Sec. 3.4.3.2) and was reported to bind the cap analogue m^{7}GTP as well as CBP80, forming an alternative nuclear cap binding complex [64], NCBP3 and its interactions were analysed in more detail. Two constructs were used for these initial studies: NCBP3 full length (residues 1-620) and NCBP3$^{1\text{-}282}$, which was previously identified as the cap-binding region [64]. The N-terminal construct NCBP3$^{1\text{-}282}$ expressed in *E. coli* showed high degradation, whereas the full length protein expressed in insect cells seemed to be more stable.

For the cap binding an affinity of 5.2 µM was determined for NCBP3$^{1\text{-}282}$ and of 4.9 µM for NCBP3 full length (Figure 3.26A). This result is comparable to the described K_D of 16 µM determined by microscale thermophoresis experiments for

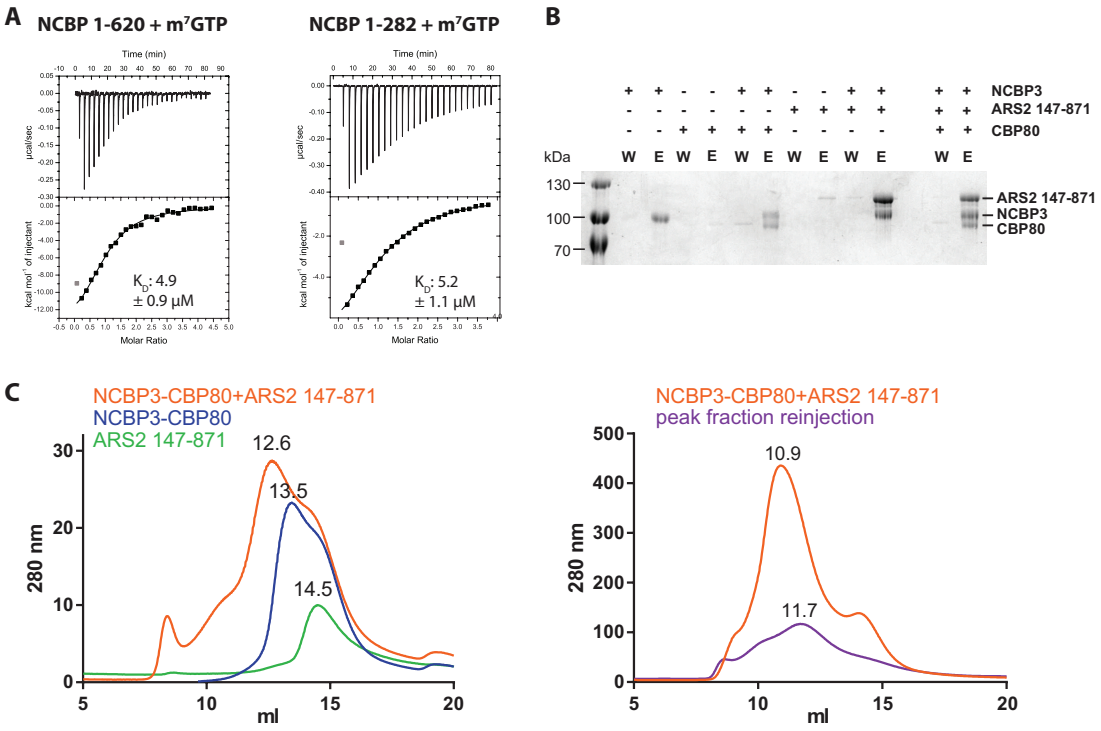


Figure 3.26 – NCBP3 binding studies

A: Determination of NCBP3's cap binding. ITC was performed with NCBP3 in the sample cell and sequential titration of m^{7}GTP. The data were plotted as described in the caption to Figure 3.3C.
B: Interaction of NCBP3 with CBP80 and ARS2. NCBP3 was immobilised on m^{7}GTP-Sepharose and incubated with CBP80 and/or ARS2. Last wash (W) and elution (E) were analysed by Coomassie-stained SDS-PAGE.
C: SEC elution profiles of (left) the formation of the NCBP3-CBP80-ARS2 complex at low concentrations and (right) the formation of the NCBP3-CBP80-ARS2 complex at high protein concentrations and reinjection of the peak fractions.

NCBP3$^{1\text{-}282}$ and salt conditions of 225 mM [64] instead of 120 mM NaCl and shows that the C-terminal residues 283-620 are not involved in cap binding.

To test the interaction of different proteins with NCBP3, NCBP3 was immobilised on m^{7}GTP-Sepharose and incubated with ARS2, CBP80, NELF-E or PHAX. The elution from the resin shows that CBP80 as well as ARS2$^{147\text{-}871}$ were retained on NCBP3 bound to the resin (Figure 3.26B), whereas no interaction of PHAX or NELF-E with NCBP3 could be detected (data not shown).

The interaction of NCBP3 with CBP80 was only observed for the full length NCBP3 and not for the N-terminal construct containing the cap binding domain (data not shown). However, the N-terminal construct NCBP3$^{1\text{-}282}$ was sufficient to bind ARS2$^{147\text{-}871}$. Furthermore, NCBP3$^{1\text{-}680}$ and NCBP3$^{1\text{-}282}$ co-eluted with GST-tagged-ARS2$^{147\text{-}871}$ in a GST pull down experiment in the absence of the cap analogue (Figure 3.27A), indicating that the interaction between NCBP3 and ARS2 is cap-independent and mediated via the N-terminal 282 amino acids of NCBP3.

Additionally, co-elution of all three proteins NCBP3, CBP80 and ARS2 was observed (Figure 3.26C left). This trimeric complex as well as the NCBP3-CBP80 and NCBP3-ARS2 complex could be reconstituted in the absence of m^{7}GTP by SEC as indicated by a shift in elution volume (Figure 3.26C). However, all these NCBP3 complexes did not seem to be very stable and disassembled when subjected to SEC at low concentrations (Figure 3.26C right).

ITC experiments were used to determine the dissociation constants and revealed a micromolar affinity of NCBP3$^{1\text{-}282}$ towards ARS2$^{147\text{-}871}$ under increased salt conditions (250 mM NaCl), which allowed to reach high enough concentrations of the proteins to perform the ITC experiment. However, the interaction was too weak to enable K_D calculation (Figure 3.27B). Due to the experimental requirement of high protein concentration at low salt concentrations no ITC experiments for the interaction of CBP80 and NCBP3 could be performed.

The observation from the EGFP-ARS2 pull down experiment that EGFP-ARS2$^{\Delta 692\text{-}743\ GSG}$ and EGFP-ARS2$^{K719A\ K722A\ K734A}$ show less enrichment of NCBP3 than ARS2 wild type (see Sec. 3.4.3.2), led to the hypothesis that the C-terminal arm of ARS2 might mediate the interaction. To test this and to identify the NCBP3 binding region on ARS2 different ARS2$^{147\text{-}871}$ constructs were tested for their NCBP3 binding *in vitro* via m^{7}GTP pull down. As shown in Figure 3.27C ARS2$^{147\text{-}871\ K719A\ K722A\ K734A}$ and ARS2$^{147\text{-}871\ \Delta 692\text{-}743\ GSG}$ did not bind to the m^{7}GTP-Sepharose immobilised NCBP3. This result was also confirmed by ITC, in which no binding of the two constructs to NCBP3$^{1\text{-}282}$ could be measured (Figure 3.27B right, data not shown for ARS2$^{147\text{-}871\ \Delta 692\text{-}743\ GSG}$). Furthermore, GST-ARS2$^{634\text{-}871}$ was sufficient to pull down NCBP3 in a GST pull down experiment, whereas GST-ARS2$^{634\text{-}871\ K719A\ K722A\ K734A}$ did not interact with NCBP3 (Figure 3.27A). These results together with the mass spectrometric analysis of the different ARS2 constructs indicate that the basic patch on the C-terminal arm of ARS2 mediates the interaction with NCBP3.

Interestingly, the same mutations also abolished the interaction with FARB (see Sec. 3.4.2.2), resulting in the question if FLASH and NCBP3 might compete for ARS2 binding. To investigate this the GST-FLASH$^{903\text{-}943}$-ARS2$^{147\text{-}871}$ complex was immobilised on a Glutathione Sepharose resin and titrated with different amounts of NCBP3 (Fig-

ure 3.27D). Analysis of the flow through and elution of each sample showed that with increasing amounts of NCBP3 less ARS2$^{147\text{-}871}$ was retained on the resin with GST-FLASH$^{903\text{-}943}$ and more ARS2$^{147\text{-}871}$ was in the flow through together with NCBP3. Thus, NCBP3 is out competing FLASH from ARS2, illustrating that only FLASH or NCBP3 can be bound to ARS2 at a time.

Assuming that NCBP3 binds in the same fashion to ARS2 as FLASH, sequence analysis identified two acidic motifs similar to that of FARB (DELEEGEIRSDSE) within the first 282 residues of NCBP3 (41-EVEEGELEI-49, 213-DEAEEGEVEDENS-225). However, mutations of these residues to alanine in NCBP3$^{1\text{-}282}$ did not abolish the binding of ARS2 in a m^{7}GTP-Sepharose pull down experiment (data not shown).

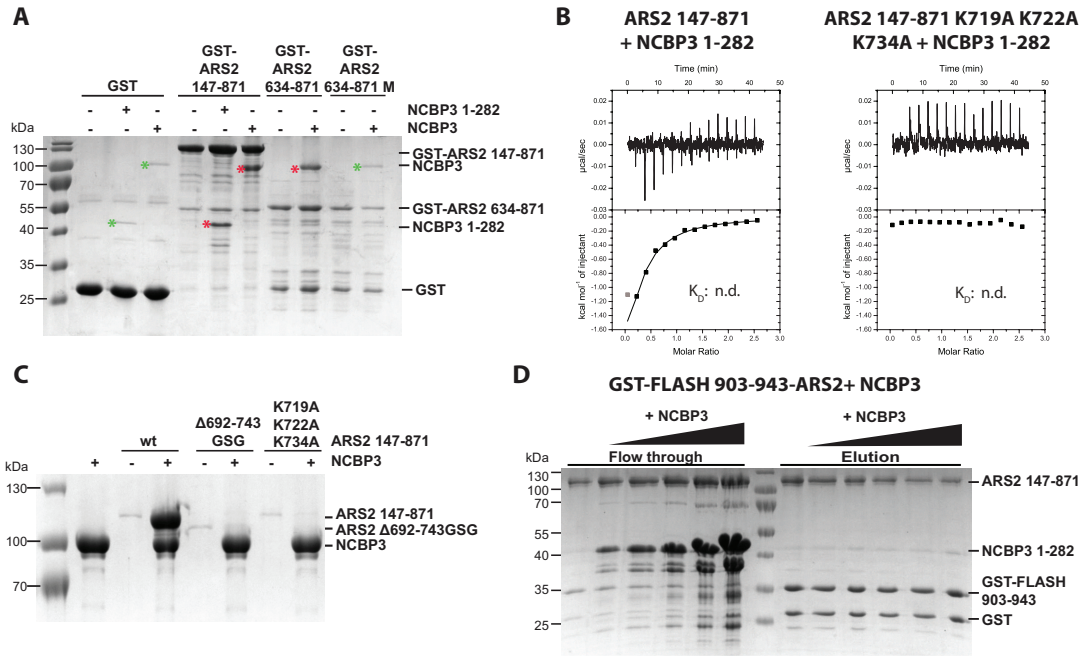


Figure 3.27 – NCBP3-ARS2 binding studies

A: Narrowing down the ARS2-NCBP3 interacting regions. Different GST-tagged ARS2 constructs were immobilised on Glutathione resin and incubated with NCBP3 full length or NCBP3$^{1\text{-}282}$. Elutions were analysed by Coomassie-stained SDS-PAGE. Red stars indicate the ARS2 bound NCBP3 constructs and green stars the unspecific binding also observed for GST alone.
B: ITC of NCBP3$^{1\text{-}282}$ with ARS2$^{147\text{-}871}$constructs. ARS2$^{147\text{-}871}$in the sample cell was titrated with NCBP3$^{1\text{-}282}$. The data were plotted as described in the caption to Figure 3.3C. Due to the weak interaction a precise K_D could not be determined.
C: Mapping the interaction site of NCBP3 on ARS2. NCBP3 was immobilised on m^{7}GTP-Sepharose and incubated with different ARS2$^{147\text{-}871}$constructs. The elutions (E) were analysed by Coomassie-stained SDS-PAGE.
D: Competition between NCBP3 and FLASH for ARS2 binding. The GST-FLASH$^{903\text{-}943}$-ARS2$^{147\text{-}871}$ complex was preformed on Glutathione Sepharose and incubated with different amounts of NCBP3$^{1\text{-}282}$. The flow through and the elution fractions were analysed by SDS-PAGE followed by Coomassie staining.

Observing the overlap of the NCBP3 and FLASH binding sites on ARS2 togther with the weak stability of the different NCBP3 complexes and the 50 times weaker affinity for the cap analog of NCBP3 compared to CBC (NCBP3 K_D~5 µM; CBC K_D~0.1 µM) the question if NCBP3 might be part of a CBC-ARS2 complex arose. Remarkably, Gebhardt *et al.* identified not only CBP80 and ARS2 within their pull down experiment but also CBP20 as NCBP3 associated protein [64]. Therefore, the compatibility of CBP20 and NCBP3 binding to CBP80 was first tested by pull down and SEC in the absence of m^{7}GTP. NCBP3 co-eluted in both experiments with CBC (Figure 3.28B, pull down data not shown), but not with CBP20 alone (data not shown). The SEC showed that a stable stoichiometric complex containing NCBP3, CBP80 and CBP20 could be formed (Figure 3.28B) unlike the NCBP3-CBP80 complex, which seemed to disassembly partially while SEC. Addition of ARS2$^{147\text{-}871}$ to this ternary complex led to a further shift in the elution volume and co-elution of all four proteins (Figure 3.28A, C). In general the CBC-NCBP3 and CBC-NCBP3-ARS2 complex seemed to form a more stable complex on the SEC compared to the NCBP3-CBP80, NCBP3-ARS2 and NCBP3-ARS2-CBP80 complex. The stabilisation of the NCBP3 complexes might be mediated by CBP20, when bound to CBP80, and/or

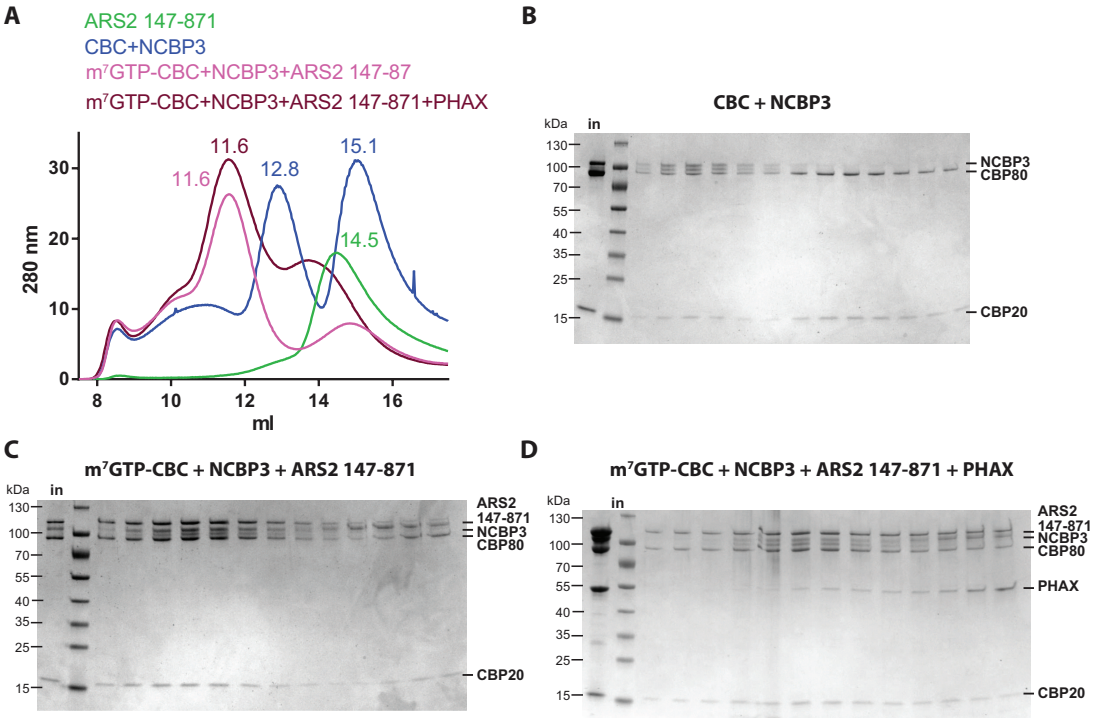


Figure 3.28 – Reconstitution of the CBC-NCBP3 and CBC-ARS2-NCBP3 complex

A: To test the interaction of CBC with NCBP3 different proteins and protein mixtures (ARS2$^{147\text{-}871}$, CBC+NCBP3, m^{7}GTP-CBC+NCBP3+ARS2$^{147\text{-}871}$ and m^{7}GTP-CBC+NCBP3+ARS2$^{147\text{-}871}$+PHAX) were injected to SEC and their elution profiles overlaid.
B-D: Analysis of the protein containing fractions of SECs (**A**) by Coomassie-stained SDS-PAGE.

a conformational change of CBP80 upon CBP20 binding. Furthermore, within CBC-NCBP3-ARS2, ARS2 is presumably interacting with CBP80 and NCBP3, thus linking the two proteins and enhancing the stability of the complex.

To broaden the studies also the compatibility of CBC-NCBP3 and PHAX was analysed. Initial experiments show that no CBC-NCBP3-PHAX or CBC-ARS2-NCBP3-PHAX complex can be formed (Figure 3.28D, data for CBC-NCBP3-PHAX not shown), suggesting that the binding of NCBP3 and PHAX to CBC is incompatible.

4 DISCUSSION

4.1 ARS2 - A PROTEIN BINDING SURFACE FACILITATING RNA SORTING AND PROCESSING

4.1.1 *Human ARS2 Differs from its Plant Homolog SERRATE*

The structure of human ARS2 combined with a systematic analysis of its RNA and protein binding provides valuable information for a better understanding of its function. The human ARS2 compared to its homolog SERRATE from *A. thaliana* [129] contains additional helices in its mid domain as well as an inserted RRM domain. Although ARS2 shows high conservation and three out of the four residues involved in zinc binding in SERRATE are present, the C-terminal arm of human ARS2 is not stabilised by a zinc ion and shows a rotation of around 90° compared to SERRATE. The loss of zinc binding can be explained by an amino acid exchange from cysteine in plants (Cys505 in SERRATE) to serine in animals (human ARS2 Ser717) (see Figure 3.19). Sequence comparison of the C-terminal domain of SERRATE and ARS2 shows an enrichment of lysines for metazoan ARS2. The observation that three of these lysines (Lys719, Lys721 and Lys734) are important for the binding of ARS2 to NCBP3 as well as to FLASH and that the C-terminal arm is not required for RNA binding indicates a different binding mechanism as well as a different function of this domain in plants and animals.

4.1.2 *The Core of Human ARS2 is Build Up by Two Fragments*

The fact that the individual fragments of ARS2 ($ARS2^{147\text{-}270}$ and $ARS2^{408\text{-}763}$) could only be expressed together and were resistant to limited proteolysis, already suggested a tight interaction of these two parts within the protein. Indeed, the structure obtained within this study shows that residues of the α-helices within the C-terminal and N-terminal arm of ARS2 are interacting and that this interaction is probably necessary for the correct folding of ARS2. However, the RRM domain, that is only present in metazoan ARS2, sits on top of the helical core and makes only few contacts to the rest of the protein, is not required for proper folding.

Comparing the sequence of the five ARS2 isoforms, only isoform 5 shows differences within the crystallised constructs of isoform 4. Isoform 5 lacks residues 196-231, which belong exactly to the two helices connecting the N-terminal arm with the helical core, while stabilising also the C-terminal arm in the solved structure. This suggests that isoform 5 differs slightly in the structural organisation of this domain and that contacts with other residues might be involved in stabilising the protein. Within the solved structure residues 232-269 include an unstructured region followed by two α-helices that are making contacts to the helical core. The expression

of ARS2$^{147\text{-}234+408\text{-}763}$ resulted in no soluble protein, suggesting that the truncation of residues 147-270 to 147-234 led to improper folding. Thus, residues 235-270 seem to be important for obtaining soluble ARS2$^{147\text{-}270+408\text{-}763}$. In isoform 5 these residues would be directly linked to the N-terminal arm and might take over the role of residues 196-231 in connecting the N-terminal arm with the helical core, while also making contacts to the C-terminal arm. Another possibility would be a complete different structure of ARS2. However, so far there are no experimental data confirming the isoform 5 as protein. Also information about the abundance of the isoforms within cell types and cell stages as well as about their functional differences are missing. As the other four isoforms differ only in one to six amino acids from each other, which lie outside of the crystallised construct, their structure will not be affected.

4.1.3 *ARS2 Contains Many Flexible, Unstructured Regions*

The structure of human ARS2 revealed many flexible regions including the N- and C-terminus, the mid domain residues 271-408 and two smaller loop regions 538-554 and 567-599. Interestingly, Thr543 within loop 538-554 is supposed to be phosphorylated according to the phosphorylation data bank [85]. Although this loop is present in different lengths in all metazoan ARS2, its residues - including the threonine - are only conserved within mammals. Until now the function of this loop and phosphorylation is still unknown. The replacement of the loop by GSA or AAA led to an abnormal subnuclear localisation of ARS2 and a cell cycle arrest in G2/M phase. However, none of the point mutations showed the same localisation as the deleted loop construct and also the mutation of all residues of the loop to alanine showed no effect within the affinity purification followed by mass spectrometric analysis and cell cycle analysis. A possibility is that this loop is important for the intranuclear transport of ARS2 or involved in the release of ARS2 from the RNA transported to this distinct nuclear areas, which might be the nucleoli. To verify that the observation indicates a function of this loop and is not due to a mislocalisation of the construct further studies including protein-interaction and cell assays need to be performed.

Another conserved motif within metazoan ARS2 is the acidic helix 'EEEE' located in loop 567-599. Within the crystal this acidic helix binds to the basic surface of the RRM domain of a neighbouring molecule. In small angle X-ray scattering (SAXS) a concentration dependent behaviour of the radius of gyration R_g was observed for ARS2, whereas SEC-MALS experiments showed no oligomerisation of ARS2 (data not shown). As the SEC-MALS experiments were carried out at lower final concentration as the SAXS experiments and no information about the concentration of ARS2 within the cell is available, no conclusion about the oligomerisation of ARS2 *in vivo* can be made at this stage. In case interaction of the 'EEEE' helix with the RRM domain of another ARS2 molecule is present within the cell, this occupation of the RRM surface and/or the 'EEEE' motif could present an inhibition of ARS2 by another ARS2 molecule. However, the crystal structure presumes that it is sterically impossible for two ARS2 molecules to donate their loop 567-599 to each other and thus to form a dimer, leaving only the possibility of an oligomerisation of ARS2.

In general, flexible unstructured regions are thought to mediate protein-protein or protein-RNA interactions. Within this study the binding of ARS2 to CBC was mapped to the C-terminal residues ($ARS2^{845\text{-}871}$), while the N-terminus contains the NLS [114, 160]. Whether these two regions mediate additional interactions as well as the function of the mid domain (residues 271-408) remain elusive. RNA binding studies showed a possible involvement of the latter region in binding to pds RNA. Thus, the region could enable RNA unwinding or specific binding of ARS2 to RNAs with a distinct secondary structure.

4.1.4 *The RRM Domain of ARS2 Binds RNA*

The RRM domain of ARS2 is highly conserved in all metazoan and is also with a divergent sequence present in *S. pombe*, whereas it is absent in plants.

The RNA binding studies illustrated that the RRM domain and not the C-terminal arm as previously proposed by O'Sullivan *et al*. binds RNA with a micromolar affinity. No sequence specificity of ARS2 apart from a difference of $ARS2^{147\text{-}270+408\text{-}763}$ binding affinity to different pdsRNAs could be observed so far. This could imply an involvement of the glutamate and lysine rich predicted unstructured region comprising residues 271-407 in pdsRNA binding. Overall, ARS2 seems to associate *in vivo* preferentially with RNAs containing a stem loop like replication dependent histone pre-mRNAs, pre-miRNAs and 7SK RNA [71, 73, 129, 160]. However, recent studies questioned the specific association of ARS2 and other proteins to RNAs by showing no strong preference for distinct classes of capped RNAs [66]. Thus, the role of the RRM domain and ARS2 in transcript class recognition remains unknown.

The highest structural similarity to ARS2 RRM domain shows the to the RRM domain of RBM7. Remarkably, RBM7 binds via its RRM domain to a peptide of ZCCHC8 within the NEXT complex [53] and the hydrophobic patch important for RBM7-ZCC5C8 interaction is also present in ARS2. This supposes that the RRM domain of ARS2 could bind both RNA and protein. To clarify, whether ARS2 can distinguish RNAs by specific sequence and/or structural elements and if the RRM serves as protein-protein interacting surface, further studies need to be performed.

4.1.5 *The C-terminal Arm of ARS2 Mediates Several Protein-Protein Interactions*

Although the RRM domain together with the C-terminal arm was previously proposed to be involved in FLASH binding [160], the binding region of the FARB peptide on ARS2 was mapped to the C-terminal arm and involves Lys719, Lys721 and Lys734. This basic region was additionally identified as binding region of NCBP3, suggesting that FLASH and NCBP3 binding to ARS2 is incompatible, which could be proven within this study *in vitro*. Until now the N-terminal region of NCBP3 (residues 1-282) was shown to interact with ARS2, whereas the C-terminal residues seem to mediate the binding to CBP80.

This gives rise to the hypothesis that also mutually exclusive ARS2 and thus CBC-ARS2 complexes are formed within RNA processing. The formation of individual CBC-ARS2-FLASH and CBC-ARS2-NCBP3 complexes is supported by the fact that

FLASH is involved in the processing of replication dependent histone mRNAs [13], whereas NCBP3 associates with mRNAs [64]. Extending this hypothesis to other processing machineries would lead to the recruitment of for example DROSHA to CBC-ARS2 complexes containing pre-miRNAs, while RNAs target for degradation would favour the recruitment of nuclear RNA decay complexes to CBC-ARS2. For snRNAs it was already shown that a distinct CBC-ARS2-PHAX complex exists [73]. Interestingly, the deletion and mutation of the basic C-terminal patch of ARS2 led to the loss of PHAX association (see Sec. 3.4.3.2). Until now a direct interaction between ARS2 and PHAX was only observed for SERRATE and not for metazoan ARS2 [66]. However, it was demonstrated that *in vivo* ARS2 stimulates the binding of PHAX to CBC [73]. Thus, the C-terminal arm might either enhance the recruitment of PHAX and/or stabilise the CBC-PHAX interaction. The observed loss of PHAX association to ARS2 (CBC-ARS2) for the two C-terminal domain mutants could be explained as following. In the case of an interaction of PHAX and ARS2, possibly via its C-terminal domain, within the CBC-ARS2-PHAX complex, the binding of other proteins to this ARS2 interaction surface would be occupied and a stable CBC-ARS2-PHAX complex be formed, whereas in the absence of the C-terminal arm or the mutation of the interacting residues the CBC-PHAX interaction is weaker and other proteins might be able to replace PHAX from the CBC-ARS2 complex.

Mutations of the C-terminally located basic patch as well as the deletion of the C-terminal arm also altered the association of ARS2 with MTR4 and ZFC3H1, proteins involved in the nuclear RNA decay. Recent studies showed that binding of ZFC3H1, which is part of the PAXT complex, and ZCCHC8/RBM7, which belong to the NEXT complex, to MTR4 is incompatible and that CBC-ARS2 is involved in both pathways [144]. Thus, the C-terminal arm of ARS2 is likely to link the RNA-CBC complex to either degradation via the NEXT or PAXT complex.

The result that also other proteins like ZC3H4 (zinc finger protein of unknown function), WDR822 (protein involved in H3-Lys4 methylation), RALY (potential hnRNP) and THRAP3 (protein involved in splicing), associate to full length ARS2 but not to the mutated and truncated ARS2 proteins, indicates that this basic patch probably links CBC-ARS2 to different processing machineries maintaining the exclusive recruitment for each step. Additionally, the affinity purification followed by mass spectrometry analysis showed an interaction between ARS2 and the export adapter ALY, which links pre-mRNA splicing to nuclear export. Whether this interaction is direct or mediated via CBC, which directly binds ALY, needs to be clarified.

Altogether these data confirm that ARS2 serves as additional interacting surface besides CBC and supports the correct sorting and processing of the nascent transcript. This might be achieved by the formation of mutually exclusive CBC-ARS2 complexes.

4.2 MUTUALLY EXCLUSIVE CBC COMPLEXES

More than 20 years ago CBC was identified as cap binding protein [94] and since then several CBC complexes have been reported to be important for the correct processing of different RNA classes. However, its role in the maturation of Pol II transcripts as well as its interactions with the different processing machineries are not well understood. This study focused on the CBC-ARS2 complex, which is involved in the maturation of different Pol II transcripts and interacts with several RNA processing machineries. Recently, a trimeric complex containing CBC-ARS2 and additionally PHAX has been shown to associate with snRNAs [73]. In the case of replication dependent histone mRNAs NELF, which binds CBC via its subunit NELF-E, as well as ARS2 seem to associate with CBC during the maturation [156]. This led to the question, if ARS2 and NELF-E are binding together to the CBC-histone mRNA complex. To better understand the individual CBC complexes and thus the individual RNA processing steps biochemical, biophysical and structural methods were used to analyse the individual interactions between CBC and ARS2, PHAX and NELF-E.

4.2.1 *Identification of Mutually Exclusive CBC Complexes*

Interestingly, the C-terminus of ARS2 (ARS2$^{845\text{-}871}$) and NELF-E (NELF-E$^{360\text{-}380}$) are mediating the binding to CBC and their affinity is increased for cap bound CBC 12- or 8-fold, respectively. For ARS2, its C-terminus seems to be the only interaction with CBC, whereas for NELF-E additional regions including the RRM domain (NELF-E$^{244\text{-}380}$) are involved and thus contribute to an increased binding affinity towards CBC. However, NELF-E$^{244\text{-}360}$, which lacks the C-terminus, was not able to bind CBC *in vitro*. For PHAX the interaction site with CBC could be narrowed down to the residues PHAX$^{103\text{-}294}$, which include most of its RBD (residues 227-308). The CBC-PHAX interaction, which in the absence of m^{7}GTP, had a similar affinity as NELF-E, showed no significant m^{7}GTP dependence.

The results obtained showed that PHAX and ARS2 binding to CBC is not compatible with NELF-E binding, suggesting a separation of different processing steps due to mutually exclusive CBC complexes. Solving the crystal structure of CBC-ARS2 and CBC-NELF-E revealed the molecular mechanism of their incompatibility. ARS2 and NELF-E are not only binding with their highly conserved C-terminus to CBC, they are also sharing the same binding site on CBC and their interactions show high similarities (see Figure 3.13). The three key interacting residues Arg362, Tyr367 and Phe380 for NELF-E and Arg854, Thy859 and Phe871 for ARS2 are identical and highly conserved. For ARS2 it was shown that Arg854 and Tyr859 are the main interacting sites and that Phe871 only contributes slightly to the interaction. Indeed, both the arginine and tyrosine residues are absolutely conserved in ARS2 and NELF-E, whereas the phenylalanine shows less conservation. For example in plants the C-terminus of ARS2, which contains a six residue extension, lacks the phenylalanine and only non-rodent mammals have an eight residue C-terminal extension of NELF-E ending in Phe380. However, not only the key interacting residues are identical, but also both peptides bound to CBC contain flexible residues (ARS2$^{862\text{-}867}$ DAPADD

and NELF-E$^{375\text{-}377}$ NLV), for which no electron density could be detected. For ARS2 these residues show high conservation, suggesting that this motif may have another functional role in case ARS2 is not bound to CBC.

Moreover, the phosphoSite database [85] contains strong evidence that ARS2 Y859 (Y864 in human isoform 1) can be phosphorylated. There is also evidence that the equivalent tyrosine in NELF-E Y367 and possible Y372 can be phosphorylated. The structures solved in this study show that these post-translational modifications would likely impede ARS2 or NELF-E binding to CBC, since the tyrosines are buried. Thus, the complex formation could be regulated by phosphorylation *in vivo*.

The incompatibility of NELF-E and PHAX binding to CBC, observed for NELF-E constructs containing additionally the RRM domain, suggests an overlapping binding site and/or steric hindrance between NELF-E and PHAX when interacting with CBC. As binding of PHAX and ARS2 to CBC are compatible, additional residues apart from the C-terminus of NELF-E must lead to the mutual exclusive binding. Interestingly, the residues of PHAX involved in CBC binding (PHAX$^{103\text{-}294}$) include most of its RRM domain (residues 227-308). Thus the RRM domain of PHAX and NELF-E seem to be in close proximity to the 5' end of the transcripts, which cap is bound by CBC. Interaction of the transcript with PHAX or NELF-E via their RRM domain could lead to the stabilisation and/or preference for one of the proteins for the RNA-CBC complex. For PHAX the binding region on CBC could only be mapped using cross-linking mass spectrometry and further investigation is needed to understand it and its incompatibility with NELF-E on molecular level.

Additionally, a novel CBC complex containing CBC-NCBP3 was identified within this study. Interestingly, NCBP3 was recently characterised to be a m^7 cap binding protein, although its affinity to the cap analogue is low. CBP80 was highlighted to interact with NCBP3, despite the fact that the data showed also an association of CBP20 with NCBP3 [64]. Earlier studies however suggested an association of NCBP3 with CBC during splicing [145]. The fact that NCBP3 binds m^{7}GTP only in the micromolar range, whereas CBC shows a 50 times higher affinity and that NCBP3 binds more stable to CBC than to CBP80 alone, raised questions about NCBP3's role as a cap binding protein *in vivo*. The result that CBC-NCBP3 was able to bind additionally ARS2 but not PHAX indicates that CBC, ARS2 and NCBP3 form an alternative CBC complex that might target the processing of specific RNA classes. This hypothesis is supported by the affinity purification mass spectrometry carried out by Gebhardt *et al.*, showing that NCBP3 is associated to CBP80, CBP20 and ARS2, whereas CBP20 shows enrichment for CBP80, ARS2 and PHAX [64]. Additionally, NCBP3 could also be identified in the CBP20 pull down, but showed no significant enrichment [64]. This could be explained by a weaker affinity of NCBP3 to CBC compared to the other proteins, a lower abundance of NCBP3 in the cell and/or the limitations of the technique. The formation of two individual CBC-ARS2 complexes containing either NCBP3 or PHAX is in line with previous observations that PHAX associates with snRNAs [158] whereas NCBP3 binds preferentially mRNAs [64] and was shown to be involved in splicing [145]. Moreover, it suggests that CBC-ARS2 bound to snRNA recruits PHAX, whereas CBC-ARS2 bound to mRNA would favor the association of NCBP3.

Furthermore, it was shown that NCBP3 and FLASH are binding both to the basic patch in the C-terminal arm of ARS2 and that their binding is incompatible. Consequently, it is likely that either a CBC-ARS2-FLASH complex or a CBC-ARS2-NCBP3 complex is formed. Thus, FLASH would associate with CBC-ARS2 complexes bound to replication dependent histone mRNAs, whereas NCBP3 interacts with CBC-ARS2 when bound to mRNA. How the different RNAs are assigned to the appropriate complex remains still to discover.

4.2.2 *Mutually Exclusive CBC Complexes Mediate RNA Processing*

In collaboration with Edouard Bertrand and Torben Heick Jensen the model of mutually exclusive CBC complexes to coordinate the fate of the RNA has recently been introduced [66]. In this study the existence of two individual CBC complexes, namely CBC-ARS2-PHAX for snRNA export and CBC-ARS2-ZC3H18 for the degradation of transcripts via the nuclear exosome was shown [66]. Additionally, another study recently showed that the exclusive binding of ZFC3H1/PABPN1 and ZCCHC8/RBM7 to MTR4 targets transcripts for exosomal degradation via the PAXT or NEXT pathway, respectively [144]. Thus, mutually exclusive binding seems to be a common mechanism within RNA processing.

Taken together all the mentioned results, the following CBC complexes seem to be mutual exclusive: CBC-NELF-E and CBC-ARS2 with CBC-ARS2-PHAX, CBC-ARS2-NCBP3 and CBC-ARS2-FLASH and CBC-ARS2-ZC3H18 with CBC-ARS2-ZC3H18-PAXT and CBC-ARS2-ZC3H18-NEXT (Figure 4.1).

But what is the biological significance and role of all these mutually exclusive CBC complexes? They could be either components of parallel but different pathways (i.e. acting on distinct RNAs) or be remodeled complexes at different time points in the biogenesis pathway of the same RNA. Considering the distinct roles of the proteins both possibilities seem plausible. In case of CBC-NELF-E, CBC-ARS2 and other CBC-ARS2 complexes a sequential rearrangement within the processing of the transcripts seems probable. NELF-E is generally thought to be associated only with Pol II pausing early after transcription initiation, whereas ARS2 is linked to later events such as 3' end processing [6, 71, 73], nuclear exosome-mediated degradation [6, 66, 144] or nuclear export [73]. Consequently, the CBC-NELF-E complexes would form first and later during transcription ARS2 would replace NELF-E from CBC (Figure 4.1).

Regarding the early co-transcriptional events, capping is thought to occur when Pol II is paused as a result of the action of the negative elongation factors DSIF and NELF. In the absence of the cap, CBC still showed *in vitro* weak affinity to NELF-E, which could be one factor helping to recruit CBC to the transcript. Once capping has occurred, CBC will bind the cap, which strengthens the CBC-NELF-E interaction. Furthermore, the ternary CBC-NELF-E-capped RNA complex can be further reinforced by NELF-E interactions, via its RRM, with the transcript, which has been proposed to be one mechanism for regulating pausing in a transcript specific manner [163]. Transcription pause release is promoted by multiple different phosphorylation events mediated by P-TEFb [127] including phosphorylation of DSIF (whereupon it becomes a

positive elongation factor), NELF (leading to its dissociation) and Pol II CTD. What exactly leads to the recruitment of P-TEFb is not well understood, but there is evidence that CBC bound to the cap structure promotes the recruitment of P-TEFb and links capping with transcription elongation [120]. Recent work demonstrated that the Integrator complex is recruited by NELF and DSIF to the paused Pol II complex and

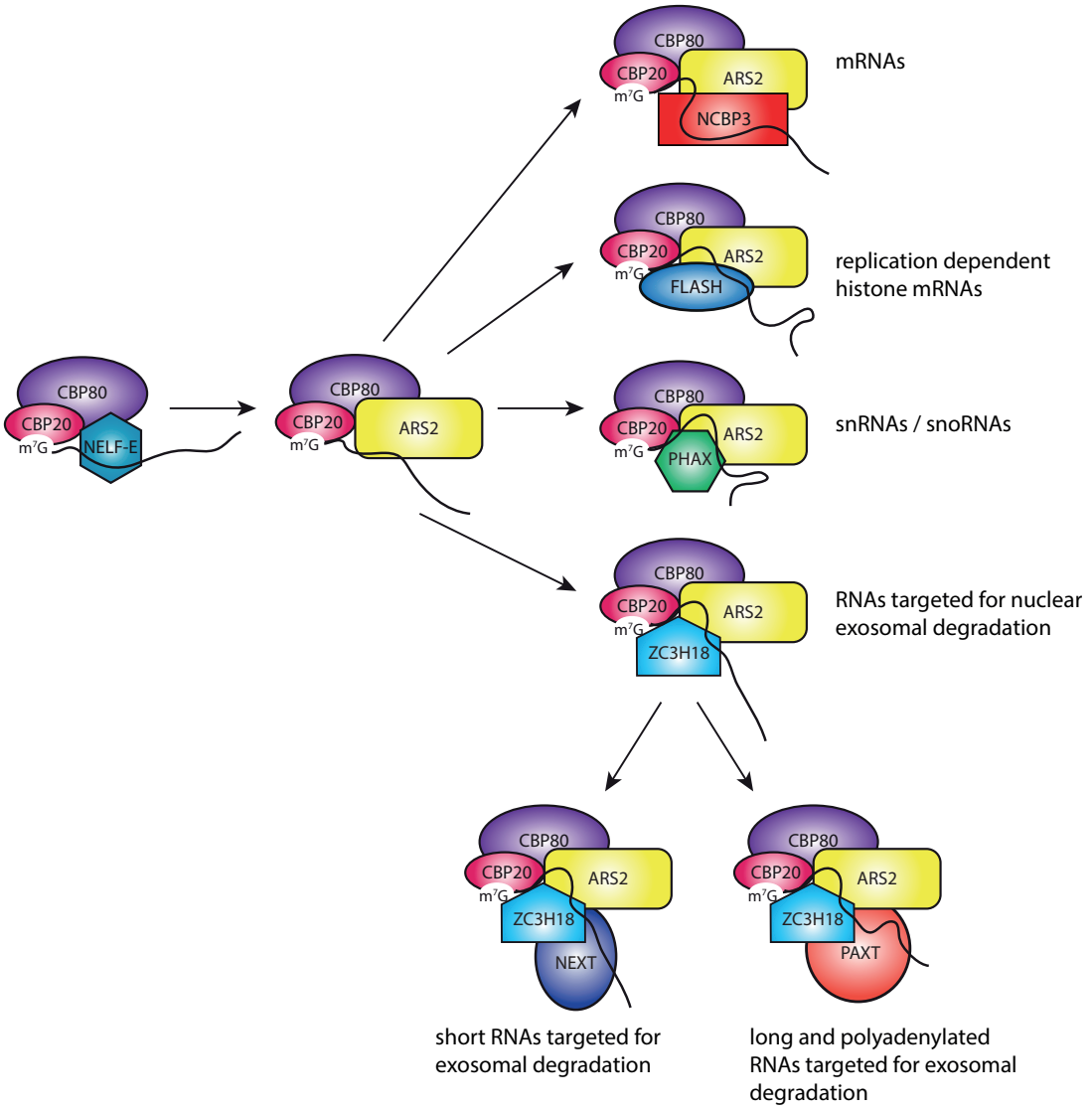


Figure 4.1 – Model of RNA processing via mutually exclusive RNA complexes

Shortly after the initiation of Pol II transcription the 5′ cap of the nascent RNA acquires a m7 cap structure, that is immediately bound by CBC, consisting of CBP20 and CBP80. Additionally, a stable complex of the transcript, CBC and NELF, which binds CBC via its subunit NELF-E, is formed. This complex could be stabilised though the binding of NELF's RRM domains to the RNA. Then at an unknown stage during transcription, NELF-E dissociated from the RNA-CBC complex and ARS2 is recruited, followed by the association of additional RNA processing factors that are specific for a distinct RNA class.

that Integrator subunits also regulate pausing [178, 196, 228]. As the precise function of the Integrator complex remains elusive, the discussion will be restricted to how the obtained results might fit into the current model of co-transcriptional processing of snRNAs [66, 228]. It was reported that NELF accumulates at the downstream regions of snRNA (and replication dependent histone mRNA) genes and is required for correct 3′ end processing. This suggests that NELF has a role in preventing aberrant polyadenylation by inhibiting recruitment of the cleavage stimulation factor. Assuming NELF is still bound to CBC at this stage, the CBC-NELF-E interaction will also prevent premature recruitment of ARS2 and PHAX. However, at some currently unknown 3′ end processing step, there must be a switch of CBC partner from NELF-E to ARS2 (Figure 4.1). This may be concomitant with NELF dissociation and/or additional phosphorylation events (e.g. as mentioned above on Y367/Y372 of NELF-E). Once the switch has happened, PHAX can be recruited to the CBC-ARS2-snRNA complex to promote its nuclear export. Alternatively, if there is some defect in 3′ end processing such as read-through, leading to extended forms of snRNA, these aberrant RNAs can be bound by the RBM7 subunit of the NEXT complex and targeted for degradation [6, 87]. Interestingly, it has recently been shown that the NEXT complex, via the bridging protein ZC3H18, interacts with CBC-ARS2, which excludes PHAX binding [66]. Thus, competition between mutually exclusive CBC complexes contributes to either the productive export of mature or degradation of incorrectly processed snRNA transcripts (Figure 4.1).

For other transcripts like mRNAs and replication dependent histone mRNAs, after the switch from NELF-E to ARS2, NCBP3 and FLASH could be recruited to the CBC-ARS2 complex, respectively. With FLASH other replication dependent histone mRNA 3′ end cleavage factors would be recruited, whereas NCBP3 seems to be part of mRNPs during splicing [145]. After correct processing or in cases of misprocessing another switch between the CBC-ARS2 bound proteins is likely to occur to either export adaptors for the nuclear export of the mature mRNA or ZC3H18 for the degradation of aberrant RNAs (Figure 4.1).

Until now no direct role of ARS2 in nuclear RNA export has been reported. For snRNAs the CBC-ARS2-PHAX complex, comprising the snRNA export adaptor PHAX, was described [158] and ARS2 seems to enhance the recruitment of PHAX to CBC [73]. Within this study an association of the mRNA export adapter ALY with ARS2 was identified via affinity purification followed by mass spectrometric analysis. Whether this interaction is direct or mediated via CBC, which is binding ALY, needs to be clarified. However, before mRNA export through the NPC the mRNP is rearranged and ALY dissociates from the complex. Already testing the compatibility of ARS2 binding to the snRNA export complex snRNA-CBC-PHAX-CRM1-Ran and the mRNA export complex mRNA-CBC-Nxf1-NXT1 will indicate, if ARS2 is co-exported with the RNA. A rearrangement of the export complexes leading to the replacement or dissociation of ARS2 from CBC, thus forming other mutual exclusive CBC complexes for RNA export, could regulate RNA export.

4.2.3 *Regulation of Mutually Exclusive CBC Complexes*

Although the existence of mutually exclusive RNA-CBC complexes has been shown and these distinct complexes are likely to regulate RNA processing, it remains unclear what coordinates the recruitment of the correct processing machinery. Taking into account that RNA processing takes part co-transcriptionally and that the Pol II transcribes a couple of kbp per minute [130], the different processes must be efficient and well-coordinated. This could be achieved by short-lived, mutually exclusive complexes that have one common factor: CBC. Depending on the RNA maturation stage additional factors such as NELF-E or ARS2 might bind to CBC and the RNA, labeling the complex for this specific maturation and/or transcription stage. As there are several different RNA classes further labels need to be added to the complex. These could be PHAX in the case of snRNA, FLASH for histone mRNA, NCBP3 for mRNA and ZC3H18 for RNA targeted for degradation. Many of the proteins involved in RNA processing contain RRM domains, which could be recruited specifically regarding the bound transcript. However, recent results showed that ARS2, PHAX and ZC3H18 bind capped transcripts without strong preferences regarding their class or maturation [66]. Furthermore, the formed complexes seemed to be very transient, presuming that exclusive transcript-CBC complexes are formed temporarily, which only results in effective processing, if all requirements are fulfilled. These requirements would include the recruitment of all factors of the distinct processing step to the RNA and specific features of the RNA that support its processing like secondary structure or sequence elements.

The assembly and disassembly of exclusive CBC complexes could further be coordinated by specific modification of proteins like phosphorylation. Phosphorylation plays an important role within transcriptional regulation of Pol II and thus also the processing could be coupled accordingly.

4.3 FUTURE PERSPECTIVES

Within this doctoral thesis, the CBC complexes CBC-ARS2, CBC-PHAX and CBC-NELF-E were studied from a molecular point of view leading to a better insight of their arrangements and the discrimination between RNA processing steps. The studies on ARS2 support its role as interacting surface within the processing of capped RNAs. However, it remains elusive how the different complexes are recruited. As distinct RNAs are processed in different ways, the nature of the RNA (e.g. length, presence of secondary structure) must determine to some extent its fate. Some of these features have already been identified. For the distinction between snRNA and mRNA it was shown that hnRNPC binds and replaces PHAX selectively from RNAs longer than 200 nts [140]. For 3' end processing different sequence and structural elements have been identified that recruit distinct processing machineries. However, the exact mechanism underlying the sorting of the individual RNA classes into the appropriate pathways is not well understood. Structural insight into protein-RNA interaction and the arrangement of the processing machineries at individual steps will help disclosing the molecular mechanism of RNA processing.

Within this study the structural characterisation was mainly based on crystallography. However, alternative methods including NMR and cryo electron microscopy can help to better understand individual domains of proteins and big complexes, respectively. Using NMR the individual interactions for example between the RRM domain of ARS2 and RNA or the C-terminal arm and FLASH and NCBP3 can be analysed. Until now although using different solubility tags no folded construct of both RRM and C-terminal arm could be purified. However, the C-terminal constructs ARS2$^{634\text{-}871}$, ARS2$^{634\text{-}763}$ and ARS2$^{677\text{-}758}$ showed weak affinity for the FARB peptide and might be a starting point for further construct optimisations for NMR studies. For the bigger complexes such as CBC-ARS2-NCBP3 or CBC-ARS2-PHAX electron microscopy can be used to obtain insights about the arrangement of the complexes. Although overall less protein might be needed for electron microscopy than for crystallography, a stable complex must be formed or the complex needs to be stabilised otherwise, for example by using chemicals to cross-link the proteins. Also the RNA itself could already lead to a stabilisation of the complex by interacting with the RRM domains of the individual proteins. Furthermore, studying the complex in context with the distinct RNAs is essential for a better understanding of RNA processing.

To help finding an RNA that could stabilise ARS2 or other components within the CBC complexes additional RNA binding studies need to be performed to identify a possible sequence or secondary structure preference. One possibility to identify a motif or specific RNAs that show high affinity for ARS2 is the affinity RNA pull down. The result would might give information about the recruitment of ARS2 to specific RNAs and will be useful for further structural studies.

Besides studies of the already described complexes, it is still unclear if ARS2 stays attached to the RNA-CBC complex during nuclear export. Testing the compatibility of ARS2 binding to the snRNA export complex CBC-PHAX-CRM1-Ran could give information about a possible export of ARS2 within this complex.

Moreover, the identification of NCBP3 as part of a CBC-ARS2 complex needs further investigations including the mapping of the ARS2 and CBP80 binding site on NCBP3, studies of its RRM domain regarding RNA and cap binding as well as its interactions with other proteins, to better define its role within RNA processing.

All these studies of the individual proteins as well as the whole complexes will help to shed light on how the different Pol II transcripts are processed and how the processing and sorting is coordinated.

A

SUPPLEMENTARY MATERIAL AND METHODS

A.1 COMPOSITION OF SOLUTIONS

The composition of buffers and media used in this study are listed Table A.1 and Table A.2 respectively.

Table A.1 – Buffer

BUFFER	COMPONENTS
Buffer A pH 7.8	120 mM NaCl
	2 mM TCEP
	20 mM HEPES
Buffer B pH 7.8	800 mM NaCl
	2 mM TCEP
	20 mM HEPES
CC1 pH 5.8	100 mM RbCl
	50 mM $MgCl_2$
	10 mM potassium acetate
	10 mM $CaCl_2$
	15% (v/v) glycerol
CC2 pH 6.5	10 mM RbCl
	75 mM $CaCl_2$
	10 mM MOPS
	15% (v/v) glycerol
HEK 293 Lysis Buffer	20 mM HEPES pH 7.4
	120 mM NaCl
	0.1% (v/v) Triton X-100
Labeling Buffer	PBS pH 7.4
	0.5% (w/v) BSA
	0.5% (v/v) Tween-20
Lysis Buffer pH 7.8	120 mM NaCl
	10% (v/v) glycerol
	5 mM ß-mercaptoethanol
	50 mM HEPES

Table A.1 – Continued on next page

Table A.1 – Buffer (continued)

BUFFER	COMPONENTS
PBS pH 7.4	137 mM NaCl
	2.7 mM KCl
	10 mM Na_2 HPO_4
	1.8 mM KH_2PO_4
RNA Binding Buffer pH 7.8	115 mM NaCl
	1 mM $MgCl_2$
	2 mM TCEP
	7.5% (v/v) glycerol
	20 mM HEPES
SDS Running Buffer pH 8.3	25 mM Tris
	192 mM glycine
	0.1% (w/v) SDS
4x SDS Sample Buffer	0.1% (w/v) bromophenol blue
	25% (v/v) glycerol
	5% (w/v) SDS
	10% (v/v) β-mercaptoethanol
	250 mM Tris pH 6.8
TAE Buffer pH 8.0	40 mM Tris
	1 mM EDTA
	20 mM acetic acid
Tris-EDTA Buffer pH 8.0	1 M Tris
	0.1 M EDTA

Table A.2 – Media

MEDIUM	COMPONENTS
LB agar	1.5% (w/v) agar in LB
High Five insect cells medium	Express Five SFM (Gibco)
	50 μg/ml penicillin-streptomycin (Gibco)
	18 mM L-Glutamine (Gibco)
HEK293T medium	high glucose Dulbecco's Modified Eagle Medium (DMEM) (Gibco)
	10% (v/v) FBS (Sigma)
M9 medium pH 7.2	33.7 mM Na_2 HPO_4
	22 mM KH_2PO_4
	8.55 mM NaCl
	9.35 mM NH_4Cl

A.2 INSTRUMENTS

Instruments used in this study are listed in Table A.3.

Table A.3 – Instruments

INSTRUMENT	TYPE	MANUFACTURER
Agarose gel chamber	Horizontal Apparatus	Bio-Rad
Centrifuges	Avanti J-26-XP	Beckman Coulter
	Centrifuge 5424	Eppendorf
	Centrifuge 5804	Eppendorf
	Centrifuge 5427 R	Eppendorf
	SpeedVac Concentrator	Savant
Chromatography columns	Superdex 200 10/300 GL	GE Healthcare
	Superdex 200 PC 3.2/30	GE Healthcare
	Superdex 75 10/300 GL	GE Healthcare
	Superose 6	GE Healthcare
	Superose 6 increase 10/300	GE Healthcare
	Superdex peptide PC3.2/30	GE Healthcare
	HiTrap Heparin HP 5 ml	GE Healthcare
	HiTrap Q FF 5 ml	GE Healthcare
	HiTrap Q HP 5 ml	GE Healthcare
	HiTrap SP HP 5 ml	GE Healthcare
Chromatography systems	Äktra Prime	GE Healthcare
	Äktra Purifier	GE Healthcare
	Äktra	GE Healthcare
Concentrator	Amicon Ultra	Merck Millipore
Flow cytometer	BD LSRFortessa	BD Bioscience
Formulator	Formulator	Formulatrix
Electrophoresis Chamber	Vertical electrophoresis cell	Bio-Rad
Gel documentation system	Gel Doc XR+	Bio-Rad

Table A.3 – Continued on next page

Table A.3 – Instruments (continued)

INSTRUMENT	TYPE	MANUFACTURER
Isothermal titration calorimeter	MicroCal iTC200 System	Malvern
Microplate reader	CLARIOstar	BMG LABTECH
Microscope	TCS SP5	Leica
NanoDrop spectrophotometer	ND-1000, ND-2000	Thermo Scientific
Spectrophotometer	Ultrospec 1100 pro UV/Visible	Amersham Pharmacia
pH meter	Microprocessor pH	WTW
Rotors	JA25.50	Beckmann Coulter
	JA8.100	Beckmann Coulter
Sonicator	Ultrasonic Liquid Processors	Misonix
Thermal block	ThermoMixer	Eppendorf
Thermocycler	T3000	Biometra
	T3	Biometra
Vortex mixer	MS2 Minishaker	IKA
Water bath	TW8 Water Bath	JULABO

A.3 KITS

Kits used in this study are listed in Table A.4.

Table A.4 – Kits

KIT	MANUFACTURER
Additive Screen HT	Hampton Research
Bradford Protein Assay	Bio-Rad
Gel filtration LMW calibration kit	GE Healthcare
In-Fusion HD Cloning Kit	Clontech
QIAprep Spin Miniprep Kit	Qiagen
QIAquick Gel Extraction Kit	Qiagen
QIAquick PCR Purification Kit	Qiagen
NucleoSpin Plasmid Kit	Macherey-Nagel
PCR clean-up Gel extraction	Macherey-Nagel
Page Ruler Prestained Protein Ladder, 10-180 kDa	Thermo Fisher Scientific
Color Prestained Protein Standard, Broad Range	New England Biolabs
MassRuler Express Forward DNA Ladder Mix, ready-to-use	Thermo Fisher Scientific
PACT premier	Molecular Dimensions
JCSG+ Screen	Molecular Dimensions
The Classics Suite	Qiagen/NeXtal
Salt Grid	Hampton Research
Wizard I&II	Rigaku Reagents
The PEGs-I	Qiagen/NeXtal

A.4 PLASMIDS

An overview of all plasmids encoding protein constructs is given in Table A.5. Within this study the following sequences were used for the individual proteins: ARS2 (isoform-e, NP_001122326/isoform-4, Q9BXP5-4, 871 residues); NELF-E (NP_002895.3); PHAX (NP_115553.2); NCBP3 (NP_001107590.1); CBP20 (NP_031388.2); CBP80 (NP_002477.1); FLASH (NP_036247.1).

Table A.5 – Plasmids

PLASMID	PROTEIN	CONSTRUCT	COMMENT
CBP20 constructs			
pRSET A	CBP20	1-471	full length, wild type
pETM30	CBP20	1-471	full length, wild type
pETM30	CBP20	1-471 Y50A	CBP20 ARS2 binding mutant
pETM30	CBP20	1-471 Y50A Y89A	CBP20 ARS2 binding mutant
pETM30	CBP20	1-471 Y50A Y89A D107R	CBP20 ARS2 binding mutant, lower m^7GTP affinity in complex with CBP80
CBP80 constructs			
pFASTBac	CBP80	20-791	ΔNLS [138, 139]
pFASTBac	CBP80	20-791 Δ653-701 G	ΔNLS ΔCC [138, 139]
pFASTBac	CBP80	20-791 Y461A R610E H651A	CBP80 ARS2 binding mutant
pFASTBac DUAL	CBP80 + NCBP3	20-791 + 1-620	co-expression CBP80/NCBP3
ARS2 constructs			
pETM11	ARS2	1-871	full length, wild type, no expression
pFASTBac	ARS2	1-871	full length, wild type, no soluble expression
pETM11	ARS2	147-871	deletion of the predicted unstructured N-terminus, slight C-terminal degradation

Table A.5 – Continued on next page

Table A.5 – Plasmids (continued)

PLASMID	PROTEIN	CONSTRUCT	COMMENT
pFASTBac	ARS2	147-871	deletion of the predicted unstructured N-terminus, slight C-terminal degradation
pETM30	ARS2	147-871	deletion of the predicted unstructured N-terminus
pETM11	ARS2	147-845	C-terminal truncation, no CBC binding
pETM11	ARS2	147-871 N496E	
pETM11	ARS2	147-871 Δ539-554 GSA	deletion of loop 539-554
pETM11	ARS2	147-871 Δ567-599 GSGSGS	deletion of loop 567-599
pETM11	ARS2	147-871 K719A K722A K734A	mutations in the C-terminal basic patch
pETM11	ARS2	147-871 Δ692-743 GSG	deletion of the C-terminal arm
pETM11	ARS2	147-871 Δ271-408 GSGSGSGS	deletion of the predicted unstructured region 271-408
pETM11	ARS2	147-871 R426G R470G R498A R502A	mutations of the basic patch within the RRM domain
pETM11	ARS2	147-871 R424A R426G R463A R465A	mutation within the RRM domain
pETM11	ARS2	147-270	predicted structured region, N-terminal fragment for co-expression
pFASTBac	ARS2	147-270	predicted structured region, high aggregation
pETM11	ARS2	163-270	N-terminal fragment for co-expression
pETM11	ARS2	171-270	N-terminal fragment for co-expression
pETM11	ARS2	193-270	N-terminal fragment for co-expression
pETM11	ARS2	147-234	N-terminal fragment for co-expression, no soluble expression

Table A.5 – Continued on next page

Table A.5 – Plasmids (continued)

PLASMID	PROTEIN	CONSTRUCT	COMMENT
pETM11	ARS2	673-871	
pETM30	ARS2	673-871	
pETM30	ARS2	673-871 K719A K722A K734A	mutations in the C-terminal basic patch
pETM11	ARS2	673-845	C-terminal truncation, no CBC binding
pETM11	ARS2	673-763	
pETM30	ARS2	673-763	
pETM30	ARS2	673-763 K719A K722A K734A	mutations in the C-terminal basic patch
pETM11	ARS2	763-871	
pETM11	ARS2	763-845	C-terminal truncation, no CBC binding
pETM11	ARS2	811-871	
pETM11	ARS2	827-871	
pETM11	ARS2	827-871 R854A Y859A	ARS2 CBC binding mutant
pETM11	ARS2	827-871 F871A	ARS2 CBC binding mutant
pETM11	ARS2	827-871 R854A Y859A F871A	ARS2 CBC binding mutant
pETM11	ARS2	408-763	predicted structured region, no soluble expression
pFASTBac	ARS2	408-763	predicted structured region, no soluble expression
pETM11	ARS2	408-494	RRM domain, no protein expression
pETM41	ARS2	408-494	RRM domain, only expression of MBP
pETM11	ARS2	408-508	RRM domain, no protein expression
pETM41	ARS2	408-508	RRM domain, only expression of MBP
pETM11	ARS2	690-750	C-terminal arm
pETM30	ARS2	690-750	C-terminal arm

Table A.5 – Continued on next page

Table A.5 – Plasmids (continued)

PLASMID	PROTEIN	CONSTRUCT	COMMENT
pETM11 SUMO	ARS2	690-750	C-terminal arm
pETM11	ARS2	691-746	C-terminal arm
pETM30	ARS2	691-746	C-terminal arm
pETM11 SUMO	ARS2	691-746	C-terminal arm
pETM11	ARS2	677-758	C-terminal arm
pETM30	ARS2	677-758	C-terminal arm
pETM11 SUMO	ARS2	677-758	C-terminal arm
Plasmids used for co-expression with pETM11			
pDUET	ARS2	408-763	predicted structured region
pDUET	ARS2	408-763 Δ539-554 GSA	deletion of loop 539-554
pET15b	ARS2	408-871 Δ567-599 GSGSGS	deletion of loop 567-599
pET15b	ARS2	408-763 Δ539-554 GSA Δ567-599 GSGSGS	deletion of loop 539-554 and loop 567-599
pET15b	ARS2	408-871 Δ764-827	fusion of structured region to C-terminus
pET15b	ARS2	494-763	RRM domain
PHAX constructs			
pETM11	PHAX	1-394	full length, wild type
pETM11	PHAX	103-308	
pETM11	PHAX	120-308	
pETM11	PHAX	103-327	
pETM11	PHAX	103-294	minimal CBC binding construct
NELF-E constructs			
pETM11	NELF-E	1-380	full length, wild type, no soluble expression
pFASTBac HTb	NELF-E	1-380	full length, wild type
pETM11	NELF-E	244-380	deletion of the N-terminal predicted unstructured region

Table A.5 – Continued on next page

Table A.5 – Plasmids (continued)

PLASMID	PROTEIN	CONSTRUCT	COMMENT
pETM11	NELF-E	244-360	C-terminal truncation
pETM11	NELF-E	335-380	no expression
FLASH construct			
pETM30	FLASH	903-943	includes FARB sequence FLASH$^{931\text{-}943}$
NCBP3 constructs			
pETM11	NCBP3	1-620	full length, wild type, no soluble expression
pFASTBac HTb	NCBP3	1-620	slight degradation
pETM11	NCBP3	1-282	high C-terminal degradation
pETM11	NCBP3	1-282 41-49A	41-49: VEEGELEI mutated to polyA
pETM11	NCBP3	1-282 213-225A	213-225: DEAEEGEVEDENS mutated to polyA
pFASTBac DUAL	NCBP3 + CBP80	1-620 + 20-791	co-expression NCBP3/CBP80
Plasmids for cell assays			
pcDNA3.1	EGFP	EGFP + GSGGGS	GSGGGS linker used for all pcDNA3.1 EGFP constructs
pcDNA3.1	EGFP-ARS2	1-871	full length, wild type
pcDNA3.1	EGFP-ARS2	147-871	N-terminal truncation
pcDNA3.1	EGFP-ARS2	1-871 K719A K722A K734A	mutations in the C-terminal basic patch
pcDNA3.1	EGFP-ARS2	1-871 Δ692-743 GSG	deletion C-terminal arm
DNA3.1	EGFP-ARS2	1-871 Δ271-408 GSGSGSGS	deletion of the predicted unstructured region 271-408
pcDNA3.1	EGFP-ARS2	1-871 R426G R470G R498A R502A	mutations of the basic patch within the RRM domain

Table A.5 – Continued on next page

Table A.5 – Plasmids (continued)

PLASMID	PROTEIN	CONSTRUCT	COMMENT
pcDNA3.1	EGFP-ARS2	1-871 R424A R426G R463A R465A	mutation within the RRM domain
pcDNA3.1	EGFP-ARS2	147-871 Δ539-554 GSA	deletion of loop 539-554
pcDNA3.1	EGFP-ARS2	147-871 Δ539-554 AAA	deletion of loop 539-554
pcDNA3.1	EGFP-ARS2	147-871 Δ539-554 15A	mutation of loop 539-554
pcDNA3.1	EGFP-ARS2	147-871 Δ567-599 GSGSGS	deletion of loop 567-599
pcDNA3.1	EGFP-ARS2	147-871 T543A	mutation of reported phosphorylated T543 [85]
pcDNA3.1	EGFP-ARS2	147-871 S549A L550A P551A S552A	549-552 AAAA
pcDNA5	mCherry		(Bertrand lab)
pcDNA5	mCherry-Nop58	Nop58	nucleoli marker (Bertrand lab)
pcDNA5	mCherry-coilin	coilin	nucleoli marker (Bertrand lab)
Plasmids for yeast two-hybrid assay			
p422	CBP20	1-471	(Bertrand lab)
p422	CBP20	1-471 Y50A	CBP20 ARS2 binding mutant
pAS	CBP80	1-791	(Bertrand lab)
pAS	CBP80	1-791 Y461E R610E H651A	CBP280 ARS2 binding mutant
pACT II	ARS2	1-871	(Bertrand lab)
pACT II	ARS2	1-871 R854A Y859A R871A	ARS2 CBC binding mutant
pACT II	Alix	Alix	negative control (Bertrand lab)

B

SUPPLEMENTARY RESULTS

B.1 CBC-ARS2 INTERACTIONS

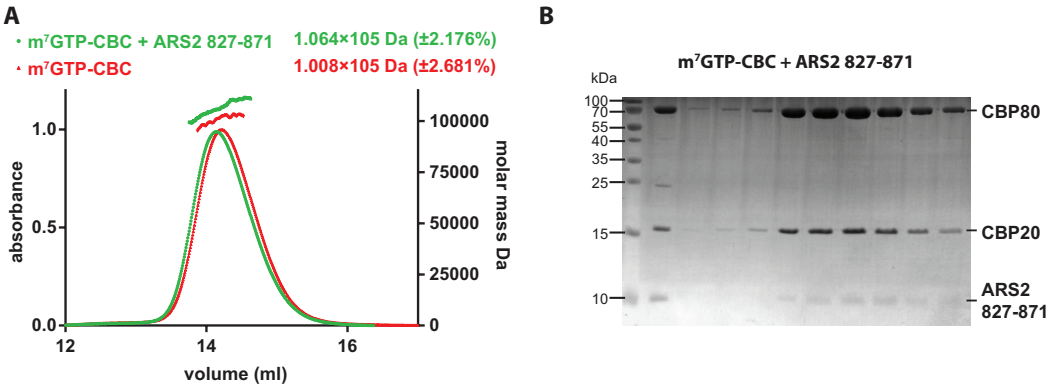


Figure B.1 – Analysis of the CBC-ARS2$^{827\text{-}871}$ complex

A: SEC-MALS of CBC and CBC together with ARS2 in the presence of m^{7}GTP were performed and revealed an 1:1 binding ratio.
B: Reconstitution of the m^{7}GTP-CBC-ARS2$^{827\text{-}871}$ complex. CBC was mixed with molar excess of ARS2$^{827\text{-}871}$and m^{7}GTP and subjected to SEC. The protein containing elution fractions were analysed by Coomassie-stained SDS-PAGE.

Table B.1 – CBC-ARS2 interactions

Different ARS2 constructs were tested for CBC binding using SEC and ITC.

INTERACTION PARTNERS	OBSERVATION
CBC + ARS2 147-871	SEC: no co-elution
m^{7}GTP-CBC + ARS2 147-871	SEC: co-elution
m^{7}GTP-CBC + ARS2 147-845	SEC: no co-elution
m^{7}GTP-CBC + ARS2 147-763	SEC: no co-elution
CBC + ARS2 827-871	SEC: no co-elution ITC: K_D~12.2 ± 3.5 µM
m^{7}GTP-CBC + ARS2 827-871	ITC K_D~1.0 ±0.2 µM
m^{7}GTP-CBC + ARS2 634-871	SEC: co-elution
m^{7}GTP-CBC + ARS2 763-871	SEC: co-elution ITC: K_D~1.0 ± 0.3 µM
m^{7}GTP-CBC + ARS2 763-845	SEC: no co-elution ITC: K_D~n.d.
m^{7}GTP-CBC + ARS2 787-871	SEC: co-elution
m^{7}GTP-CBC + ARS2 811-871	SEC: co-elution
m^{7}GTP-CBC + ARS2 845-871	ITC: K_D~1.1 ± 0.16 µM
m^{7}GTP-CBC + ARS2 827-871 F871D	ITC: K_D~1.8 ± 0.26 µM
m^{7}GTP-CBC + ARS2 827-871 R854A Y859A	ITC: K_D~n.d.
m^{7}GTP-CBC + ARS2 827-871 F871D R854A Y859A	ITC: K_D~n.d.

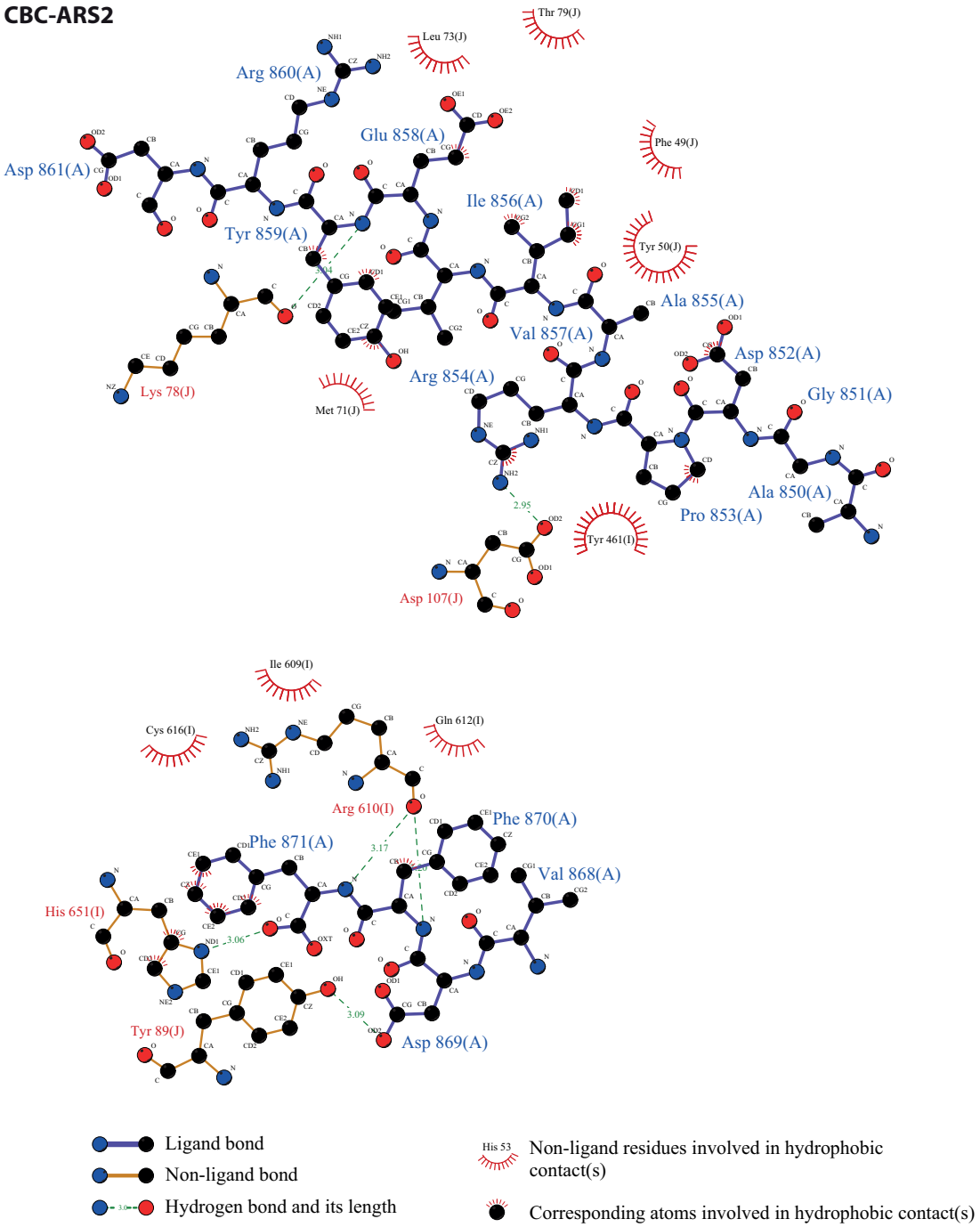
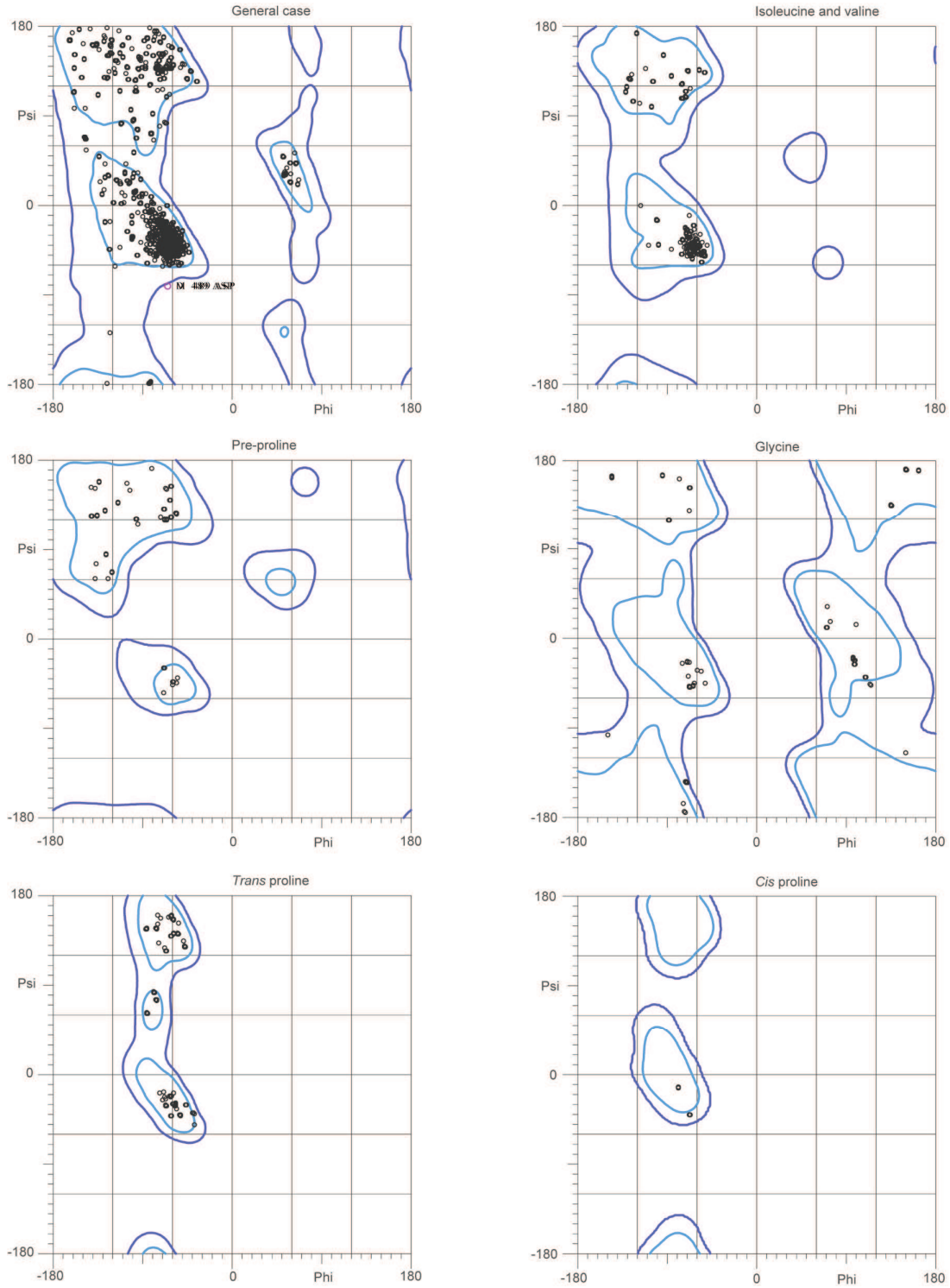


Figure B.2 – Overview of interactions within the ARS2-CBC binding site

LIGPLOT [210] was used to visualise the CBC-ARS2 peptide interactions. ARS2 bonds and residues are shown in blue and CBC bonds in orange.



97.8% (6876/7030) of all residues were in favored (98%) regions.
100.0% (7028/7030) of all residues were in allowed (>99.8%) regions.
There were 2 outliers (phi, psi): D 489 ASP (-65.5, -81.2)
M 489 ASP (-65.5, -81.3)

Figure B.3 – Ramachandran plot for m^{7}GTP-CBCΔNLS-ARS2$^{827\text{-}871}$ structure

Validation of the structure of m^{7}GTP-CBCΔNLS-ARS2$^{827\text{-}871}$ was performed with MolProbity [31].

B.2 IDENTIFICATION OF CBC-PHAX CROSS-LINKS

Table B.2 – Identified intra-protein CBC-PHAX cross-links

The CBC-PHAX complex was reconstituted *in vitro* and cross-linked with DSS. After proteolysis the purified cross-linked peptides were analysed by mass spectrometry. Cross-linked lysines are in italic and underlined.
ld (linear discriminant) confidence scores were calculated by xQuest/xProphet. The cross-links were deemed highly confident if their ld score was above 30. Cross-linked peptides marked with * were identified at least twice.

Cross-linked peptide sequence	protein1	protein2	AA 1	AA 2	ld-score
NHPQMIAVLVD***K***MIR-TDWDAGF***K***EGR	CBP8-0	CBP20	607	120	36.51
LGNRPEMNY***K***GR-TDWDAGF***K***EGR	PHAX	CBP20	216	120	35.25
KESQEHTK-SHW***K***ER*	PHAX	CBP80	162	327	34.91
HVL***K***IQK-SGDI***K***K	CBP80	CBP20	654	67	34.57
KESQEHTK-SHW***K***ER*	PHAX	CBP80	162	327	34.1
NHPQMIAVLVD***K***MIR-***K***IIMGLDK	CBP80	CBP20	607	68	33.35
KESQEHTK-SHW***K***ER*	PHAX	CBP80	162	327	32.24
KESQEHTK-SHW***K***ER*	PHAX	CBP80	162	327	31.6
TDWDAGF***K***EGR-SHW***K***ER*	CBP20	CBP80	120	327	31.03
TDWDAGF***K***EGR-SHW***K***ER*	CBP20	CBP80	120	327	30.84
TDWDAGF***K***EGR-IIGN***K***K	CBP20	PHAX	120	256	30.03
DLD***K***ELDEYMHGGKK-GDNEEQE***K***LLK	PHAX	CBP20	173	34	28.13
DLD***K***ELDEYMHGGK-GDNEEQE***K***LLKK	PHAX	CBP20	173	34	15.41
SGDIK***K***IIMGLDK-***K***DAEMDR	CBP20	CBP80	68	188	15.21

B.3 CBC-NELF-E INTERACTIONS

Table B.3 – CBC-NELF-E interactions

Different NELF-E constructs were tested for CBC binding using SEC and ITC.

INTERACTION PARTNERS	OBSERVATION
CBC + NELF-E	SEC: no co-elution
m^{7}GTP-CBC + NELF-E	SEC: co-elution ITC: K_D~0.07 ± 0.02 µM
CBC + NELF-E 244-380	SEC: no co-elution ITC: K_D~0.4 ± 0.2 µM
m^{7}GTP-CBC + NELF-E 244-380	SEC: co-elution ITC: K_D~0.05 ± 0.01 µM
m^{7}GTP-CBC + NELF-E 244-360	SEC: no co-elution ITC: K_D~n.d.
m^{7}GTP-CBC + NELF-E 354-380	ITC: K_D~3.8 ± 1.4 µM
m^{7}GTP-CBC + NELF-E 360-380	ITC: K_D~3.3 ± 1.1 µM

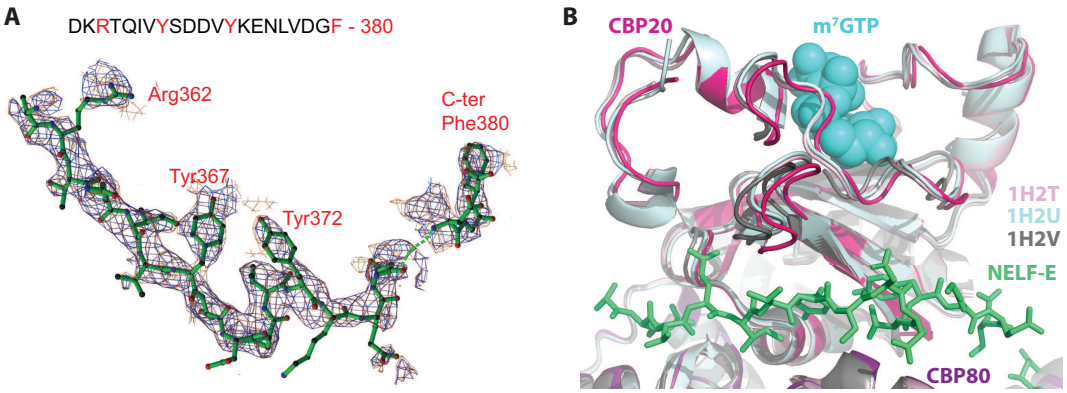


Figure B.4 – NELF-E peptide within the CBC-NELF-E crystal structure

A: Omit difference density map for NELF-E peptide contoured at 2σ. The main interacting residues are labeled and the highly conserved residues within the C-terminus of NELF-E are highlighted in red within the sequence.
B: Superposition of CBC structures (1H2V apo (gray), 1H2T (rose) and 1H2U (light blue) bound to m^{7}GpppG (cyan)) with m^{7}GTP-CBC-NELF-E$^{360\text{-}380}$ via CBP80 shows successive relative displacements of the CBP20 subunit.

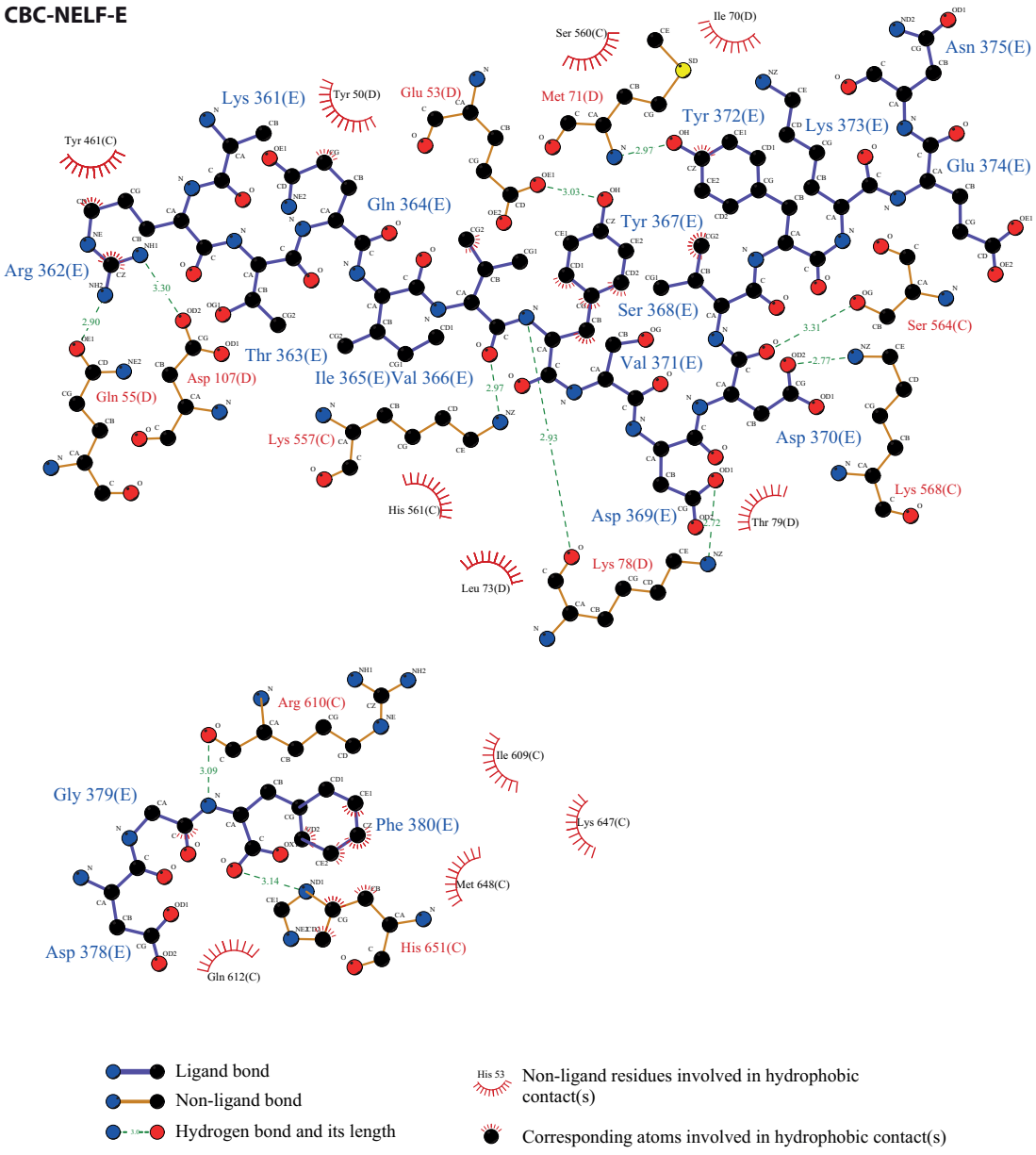
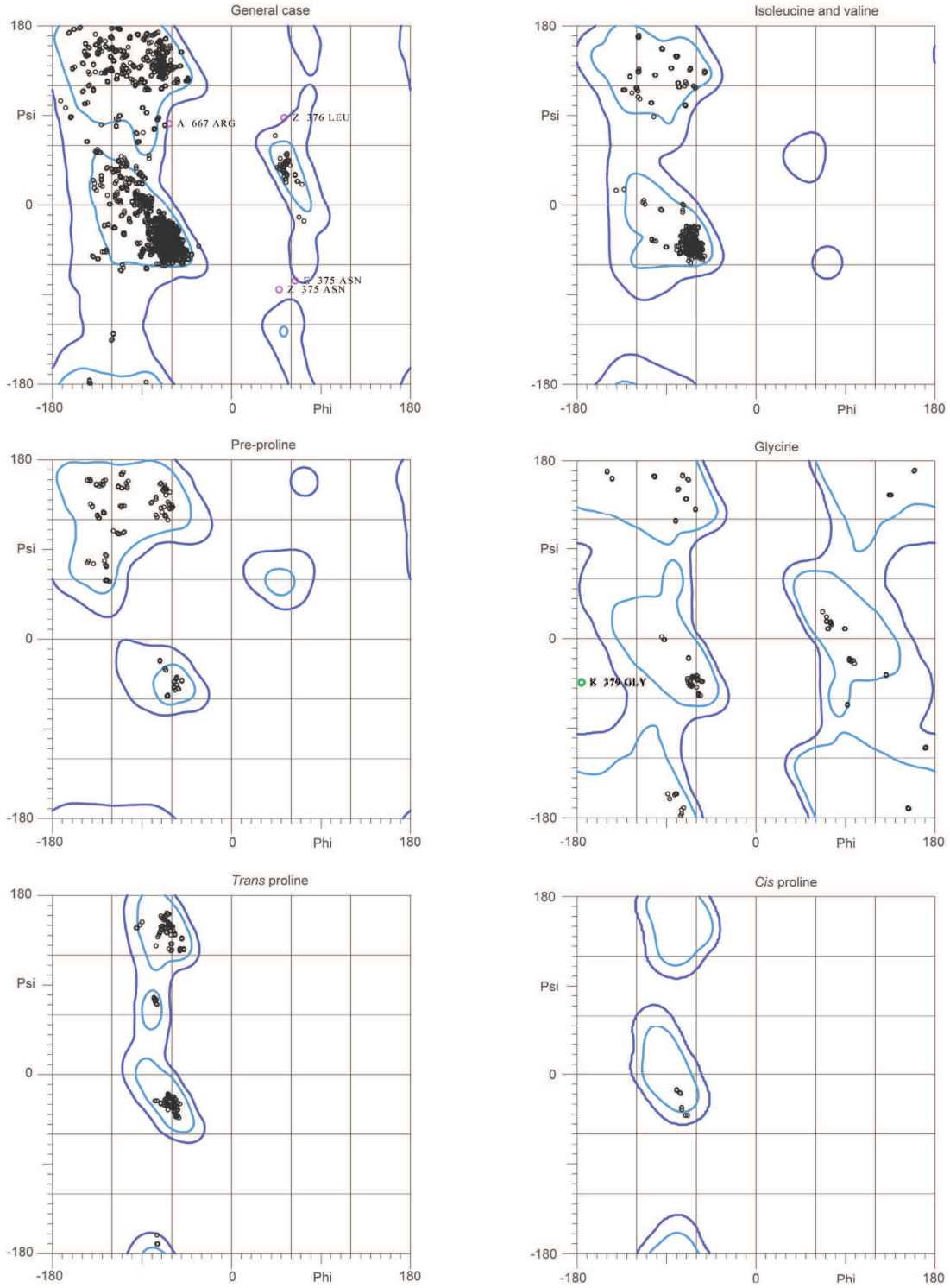


Figure B.5 – Overview of interactions within the NELF-E-CBC binding site

LIGPLOT [210] was used to visualise the CBC-NELF-E peptide interactions. NELF-E bonds and residues are shown in blue and CBC bonds in orange.



96.7% (3407/3523) of all residues were in favored (98%) regions.
99.8% (3516/3523) of all residues were in allowed (>99.8%) regions.

There were 7 outliers (phi, psi):A 667 ARG (-64.0, 82.6); E 375 ASN (64.7, -76.4); E 379 GLY (-176.6, -45.1); K 379 GLY (-176.3, -43.9); Z 375 ASN (48.3, -85.3); Z 376 LEU (53.6, 88.4); Z 379 GLY (-175.1, -44.7)

Figure B.6 – Ramachandran plot for m^{7}GTP-CBCΔNLS-NELF-E$^{360\text{-}380}$ structure

Validation of the structure of m^{7}GTP-CBCΔNLS-NELF-E$^{360\text{-}380}$ was performed with MolProbity [31].

B.4 CBC-ARS2/PHAX/NELF-E INTERACTION ASSAYS

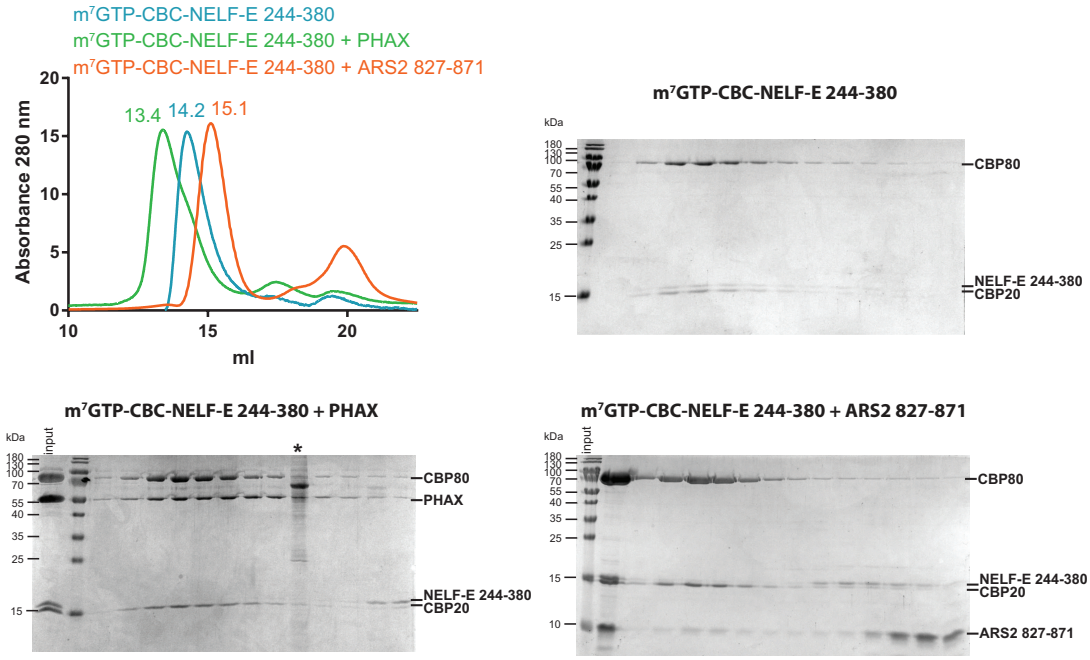


Figure B.7 – Competition of NELF-E and ARS2/PHAX for CBC-binding

The *in vitro* reconstituted m7GTP-CBC-NELF–E$^{244\text{-}380}$complex was mixed with molar excess of PHAX or ARS2$^{827\text{-}871}$ and subjected to SEC. The individual elution profiles were overlaid and the protein containing fractions analysed by Coomassie-stained SDS-PAGE. The fraction marked with * showed an additional bands, which seems to be due to a contamination of the collection tube as the elution profile shows no high absorbance for this fraction.

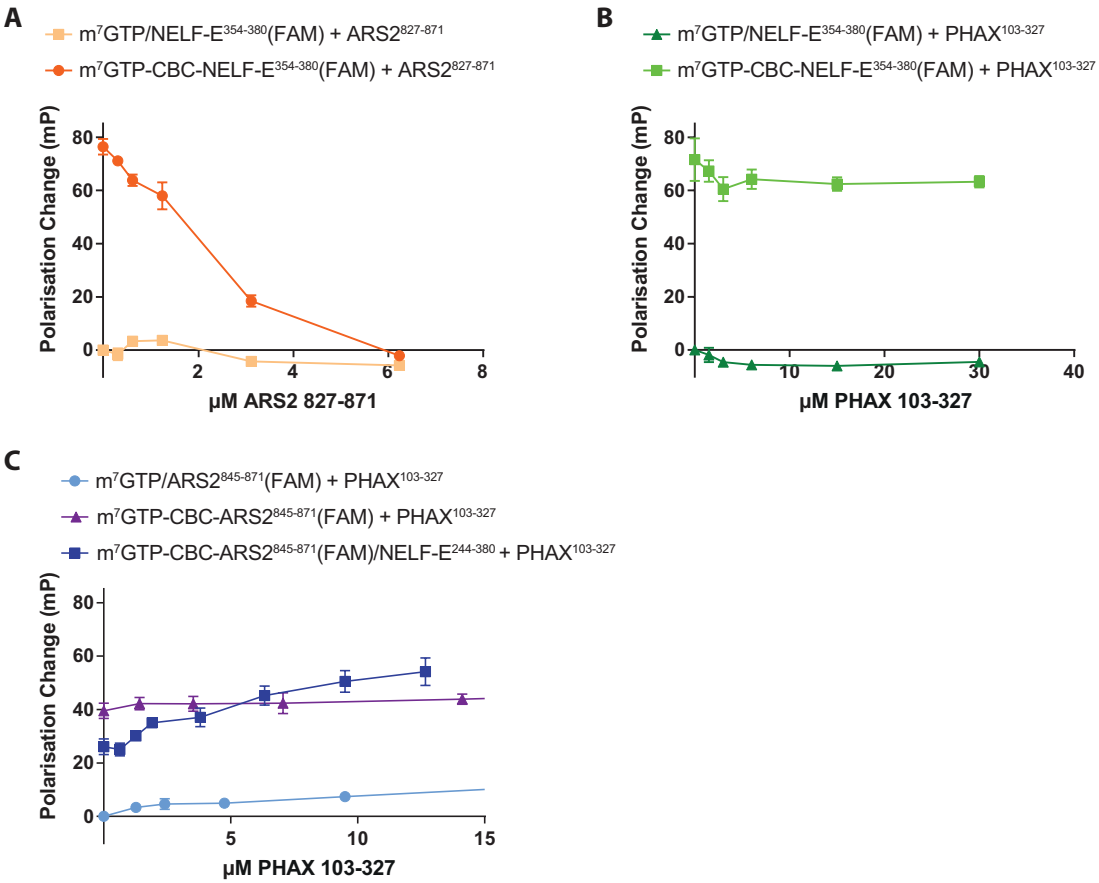


Figure B.8 – Fluorescence polarisation experiments including controls

To exclude interactions between ARS2$^{827\text{-}871}$(FAM) and PHAX or NELF-E and NELF-E$^{354\text{-}380}$(FAM) and ARS2 or PHAX control experiments without CBC were performed. To validate the competition between PHAX and NELF-E, PHAX was titrated to a preformed m^{7}GTP-CBC-ARS2$^{845\text{-}871}$(FAM) complex.
A: ARS2$^{827\text{-}871}$ titrated to m^{7}GTP/NELF-E$^{354\text{-}380}$(FAM) and
to m^{7}GTP-CBC-m^{7}GTP-NELF-E$^{354\text{-}380}$(FAM).
B: PHAX$^{103\text{-}327}$ titrated to m^{7}GTP/NELF-E$^{354\text{-}380}$(FAM) and
to m^{7}GTP-CBC-m^{7}GTP-NELF-E$^{354\text{-}380}$(FAM).
C: PHAX$^{103\text{-}327}$ titrated to m^{7}GTP/ARS2$^{845\text{-}871}$(FAM),
to m^{7}GTP-CBC-m^{7}GTP-ARS2$^{845\text{-}871}$(FAM) and
to m^{7}GTP-CBC-ARS2$^{845\text{-}871}$(FAM)/NELF-E$^{244\text{-}380}$.

B.5 STRUCTURAL ANALYSIS OF ARS2

Table B.4 – Diffraction data statistics for human ARS2

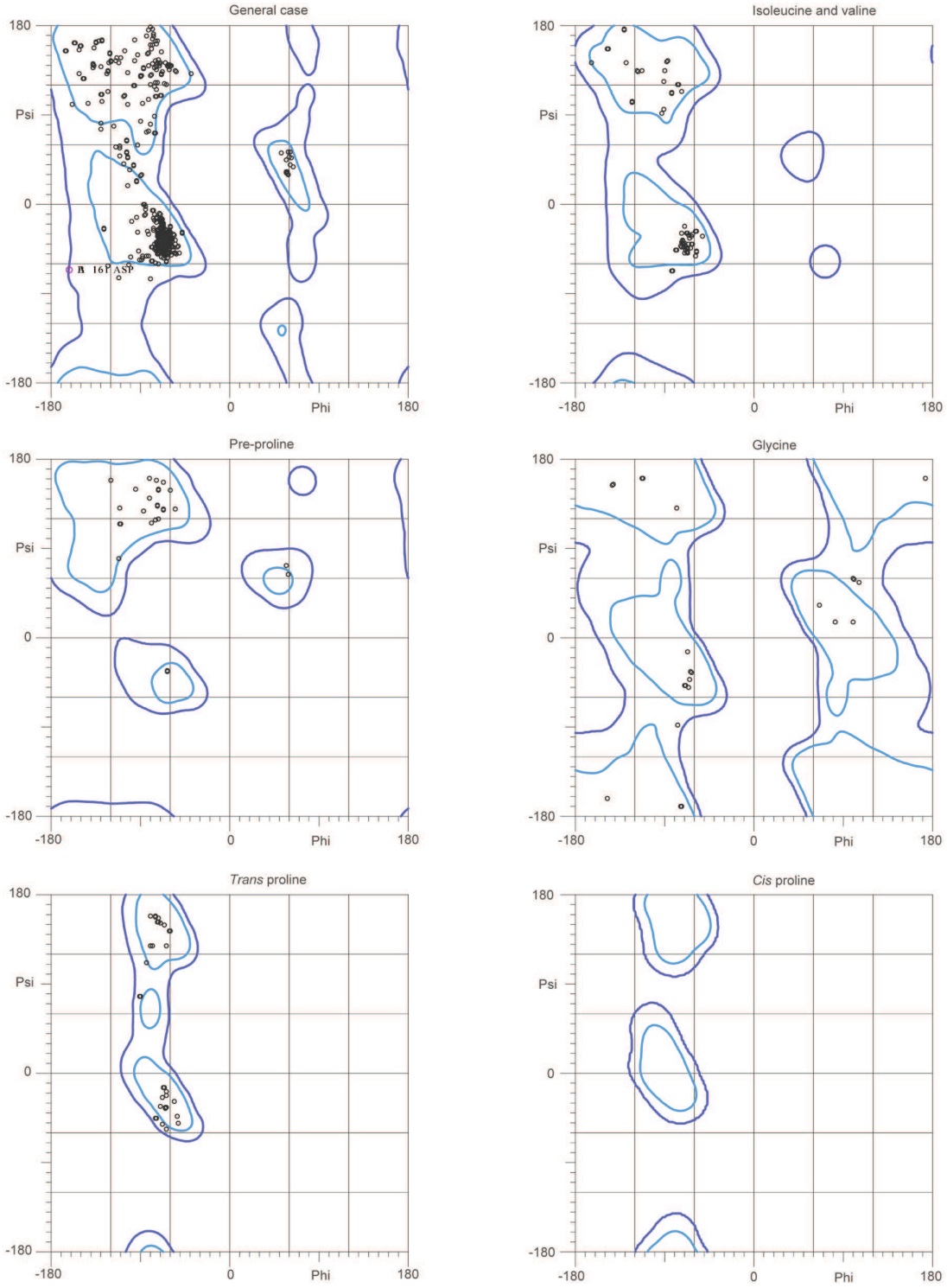
CRYSTAL	147-763 trypsinised native
Diffraction Data	
Beamline	ID29
Wavelength (Å)	0.984
Space group	$P6_4$
Cell dimensions (Å)	a=b=117.7 c=147.9
Cell angles (°)	α=β= 90 γ=120
Resolution range of data (last shell) (Å)	50.0-6.98 (7.17-6.98)
Completeness (last shell) (%)	99.3 (99.4)
R-sym (last shell) (%)	5.8 (102.6)
I/σI (last shell)	22.85 (2.00)
CC(1/2) (last shell) (%)	100.0 (85.3)
Redundancy (last shell)	11.1 (10.3)

Table B.5 – Diffraction data and refinement statistics for human ARS2 I

CRYSTAL	147-270 +408-763 native	171-270 +408-763 + 0.1 M FARB native	171-270 +408-763 + 0.1 M FARB native
Diffraction Data			
Beamline	ID30A1	ID29	ID29
Wavelength (Å)	0.966	0.984	0.984
Space group	$P3_121$	$P6_522$	$P6_5$
Cell dimensions (Å)	a=b=136.64 c=158.79	a=b=90.92 c=267.02	a=b=105.21 c=267.04
Cell angles (°)	α=β=90 γ=120	α=β=90 γ=120	α=β=90 γ=120
Resolution range of data (last shell) (Å)	50.0-3.7 (3.8-3.7)	50.0-3.22 (3.3-3.22)	50.0-3.48 (3.57-3.48)
Completeness (last shell) (%)	99.7 (98.2)	97.0 (99.9)	97.9 (86.6)
R-sym (last shell) (%)	8.1 (115.8)	25.9 (167.5)	11.4 (161.6)
I/σI (last shell)	12.11 (1.44)	10.15 (1.42)	8.71 (0.80)
CC(1/2) (last shell) (%)	100 (76.3)	99.6 (54.4)	99.7 (2.3)
Redundancy (last shell)	6.33 (6.16)	10.13 (9.13)	3.12 (2.98)
Refinement			
Reflections used in refinement work (free)	17679 (979)	10435 (516)	20048 (922)
R-work (last shell)	0.282 (0.429)	0.285 (0.398)	0.276 (0.452)
R-free (last shell)	0.311 (0.437)	0.323 (0.424)	0.301 (0.517)
Number of non-hydrogen atoms	7194	3433	6530
Geometry and B-factors			
RMSD (bonds) (Å)	0.007	0.007	0.007
RMSD (angles) (°)	1.03	0.990	0.963
Ramachandran favoured (%)	91.2	95.1	91.4
Ramachandran outliers (%)	1.4	0.5	0.52
Clash score	0.97	0.44	2.52
MolProbity score	1.55	1.35	1.73
Average B-factor	203.3	81.7	161.2

Table B.6 – Diffraction data and refinement statistics for human ARS2 II

CRYSTAL	171-270 +408-763 seleno-Met	147-270 +408-763 Δ569-599 native	147-270 +408-763 Δ539-555 native
Diffraction Data			
Beamline	ID30A1	ID30A1	ID30A1
Wavelength (Å)	0.966	0.966	0.966
Space group	$P6_522$	$C222_1$	$P3_121$
Cell dimensions (Å)	a=b=90.29 c=265.72	a=85.51 b=148.27 c=235.66	a=b=139.58 c=156.16
Cell angles (°)	α=β=90 γ=120	α=β=γ=90	α=β=90 γ=120
Resolution range of data (last shell) (Å)	50.0-3.30 (3.40-3.30)	50.0-3.37 (3.46-3.37)	45.04-3.50 (3.60-3.50)
Completeness (last shell) (%)	98.6 (99.9)	94.3 (67.1)	99.9 (99.8)
R-sym (last shell) (%)	32.4 (245)	9.8 (67.4)	9.4 (258.4)
I/σI (last shell)	10.75 (1.45)	12.62 (1.97)	13.75 (0.77)
CC(1/2) (last shell) (%)	99.7 (67.6)	99.8 (53.7)	99.9 (61.8)
Redundancy (last shell)	19.82 (17.78)	4.12 (3.82)	9.93 (9.60)
Refinement			
Reflections used in refinement work (free)	9723 (478)	19276 (1017)	21538 (1112)
R-work (last shell)	0.274 (0.307)	0.257 (0.383)	0.312 (0.466)
R-free (last shell)	0.309 (0.373)	0.300 (0.448)	0.355 (0.412)
Number of non-hydrogen atoms	3447	7146	7194
Geometry and B-factors			
RMSD (bonds) (Å)	0.007	0.007	0.008
RMSD (angles) (°)	0.968	0.949	1.016
Ramachandran favoured (%)	92.7	93.9	91.4
Ramachandran outliers (%)	0.74	0.24	2.11
Clash score	0.58	0.91	2.07
MolProbity score	1.35	1.40	1.65
Average B-factor	102.9	124.9	215.5



93.9% (798/850) of all residues were in favored (98%) regions.
99.8% (848/850) of all residues were in allowed (>99.8%) regions.

There were 2 outliers (phi, psi): A 161 ASP (-163.0, -66.5)
B 161 ASP (-162.6, -66.5)

Figure B.9 – Ramachandran plot for ARS2$^{147\text{-}270+408\text{-}763\Delta567\text{-}599}$ structure

Validation of the structure of ARS2$^{147\text{-}270+408\text{-}763\Delta567\text{-}599}$ was performed with MolProbity [31].

B.6 IDENTIFICATION OF ARS2 INTERACTION PARTNERS

Identified interaction partners of EGFP-ARS2 and difference detected for the enrichment of proteins for the wildtype and ARS2$^{K712A\ K722A\ K734A}$ and ARS2$^{\Delta 692\text{-}743\ GSG}$ as well as the enrichment of proteins for EGFP compared to the sample transfected with the empty plasmid are listed in Table B.7.
The interaction partners were captured by affinity chromatography (Figure B.10) followed by mass spectrometric analysis. For Limma analysis the replicates were combined and a protein was considered as significantly enriched (Table B.7), if its adjusted p-value is below 0.05 and the fold change above two. The volcano plots comparing the enrichment of EGFP-ARS2 wild type associated protein against the enrichment of the proteins for EGFP, EGFP-ARS2$^{K719A\ K722A\ K734A}$ and EGFP-ARS2$^{\Delta 692\text{-}743\ GSG}$ as well as the control EGFP-empty plasmid are shown in Sec. 3.4.3.2 Figure 3.25.

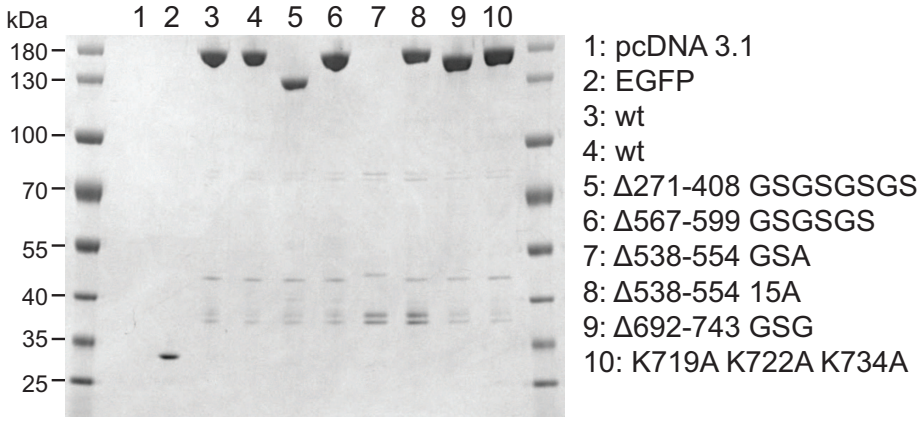


Figure B.10 – Affinity purification of EGFP-ARS2 constructs

48 h post transfection cells were harvested, lysed and the EGFP tagged ARS2 constructs purified using GFP-Trap. The elutions were analysed by SDS-PAGE followed by Coomassie staining.

Table B.7 – Identified ARS2 interaction partners

PROTEIN	log2 (fold change)	p-value	adjusted p-value
EGFP- ARS2 wt - EGFP			
NCBP3	3.707205166	2.3965E-06	0.00101931
ZC3H4	3.48016806	4.74707E-06	0.00101931
KPNA4	3.51978504	6.28858E-06	0.00101931
RBMX	2.482868451	6.82934E-06	0.00101931
KPNA2	2.406548938	7.34809E-06	0.00101931

Table B.7 – Continued on next page

Table B.7 – Identified ARS2 interaction partners (continued)

PROTEIN	log2 (fold change)	p-value	adjusted p-value
PYCR2	2.555103913	8.40889E-06	0.00101931
KPNA3	2.53445435	1.16707E-05	0.00101931
TRA2B	1.768313987	1.17915E-05	0.00101931
WDR82	2.110603166	1.24206E-05	0.00101931
THRAP3	2.08454206	1.2874E-05	0.00101931
PDCD6	2.309762297	1.35089E-05	0.00101931
PLOD3	2.774852473	1.92175E-05	0.001227944
PHAX	4.049685359	1.92329E-05	0.001227944
INPP5K	1.556435136	2.27941E-05	0.001351365
DDX5	2.227303277	2.66309E-05	0.001473578
PLOD1	2.240920311	3.98423E-05	0.002013094
TUBB4B	1.760164722	4.1232E-05	0.002013094
DDX17	2.566519356	5.07625E-05	0.002260791
LUC7L	2.67796549	5.1753E-05	0.002260791
ALY/REF	2.334103834	5.51108E-05	0.002287097
ZFC3H1	2.289790107	6.01963E-05	0.002368146
RBM14	1.522553166	6.37365E-05	0.002368146
BCKDK	2.572257584	6.56233E-05	0.002368146
PYCR1	1.901144505	7.55868E-05	0.002614044
COLGALT1	2.160954992	8.41819E-05	0.002687347
HSPA1B | HSPA1A	2.181931341	0.000102396	0.003147731
PPP2R5C	3.35292468	0.000135439	0.004014788
CBP80	3.429717745	0.000166668	0.004644559
DDX3X	1.766630902	0.000193711	0.005024384
ZCCHC8	2.263541664	0.000217439	0.005468921
SRSF10	1.600775352	0.000229266	0.005591209
FAM120A	2.219510248	0.000235774	0.005591209
MTR4	1.830769374	0.000260394	0.006003533
RALY	1.518538806	0.000287034	0.006438865
SRSF9	1.477396193	0.000355608	0.007767232
KPNA6	1.303844472	0.000489255	0.010152045
STUB1	1.27173672	0.00052102	0.010547482
THOC2	2.035342902	0.000538233	0.010636515

Table B.7 – Continued on next page

Table B.7 – Identified ARS2 interaction partners (continued)

PROTEIN	log2 (fold change)	p-value	adjusted p-value
CASC3	1.513791359	0.000581895	0.010976658
HNRNPK	1.208927669	0.000619656	0.011180744
HNRNPUL1	1.084229926	0.000766377	0.012886399
SAFB2	1.25120203	0.000792252	0.012886399
CPSF2	2.015977269	0.000793305	0.012886399
LUC7L3	1.350123837	0.000822866	0.012886399
SRSF7	1.346829717	0.000891312	0.013699794
GBAS	2.403728479	0.000945514	0.014268668
THOC1	1.660122866	0.001044681	0.015483668
SAFB	1.238422047	0.001063591	0.015487377
USP7	1.345917852	0.001146098	0.016401061
PPP2R1A	1.515670069	0.0012342	0.016811671
PPP2CA	1.110134966	0.001313657	0.017036494
HNRNPH3	1.424191363	0.001535749	0.018745166
TRA2A	1.139792449	0.001568274	0.018864742
CBP20	2.496608623	0.001614136	0.019139037
MRPS27	1.568009154	0.001790276	0.020018442
DHX30	1.012875561	0.001808895	0.020018442
RPS27	1.30546679	0.001950129	0.021297458
HNRNPA0	1.006488389	0.002106293	0.02212941
KPNA1	1.091074899	0.002148586	0.022204769
PRMT5	1.066808247	0.002462182	0.024621823
SRSF5	1.331080617	0.002527885	0.024977907
AGGF1	2.181359242	0.002633092	0.025254025
HSPA8	1.265098503	0.002647109	0.025254025
RIOK1	1.19838374	0.003215876	0.029331616
DAP3	1.421862281	0.003316398	0.029919682
WDR77	1.071060009	0.003475335	0.031016428
HNRNPF	1.230006549	0.003718166	0.03189783
AKAP8L	1.500890324	0.004372698	0.034813738
DNAJA2	1.108010879	0.004710324	0.036199715
HNRNPH2	1.243924289	0.004870933	0.037090594
PRPF38B	1.900155997	0.005435607	0.039230906

Table B.7 – Continued on next page

Table B.7 – Identified ARS2 interaction partners (continued)

PROTEIN	log2 (fold change)	p-value	adjusted p-value
FUS	1.588197445	0.005929165	0.041354681
SSBP1	1.056819114	0.006429627	0.042476399
IPO11	1.41079766	0.006448224	0.042476399
CPSF3	1.40694141	0.006825053	0.044604677
HNRNPC	1.452460114	0.007155421	0.04568461
CTNND1	1.043731251	0.007449023	0.046381391
MRPS18B	1.397996745	0.007656738	0.046493091
EIF4B	1.447075063	0.007759779	0.046671133
HSPA4	1.206122999	0.007949629	0.047469007
PPP2R2A	1.167941198	0.008254726	0.048938732
empty plasmid - EGFP			
PARK7	2.239308991	0.000356612	0.295988223
EGFP- ARS2 wt - EGFP-ARS2 K719A K722A K734A			
WDR82	2.553492114	1.64412E-06	0.001364617
NCBP3	3.286340045	8.61842E-06	0.002645791
ZC3H4	3.255334604	9.5631E-06	0.002645791
THRAP3	2.024485379	1.73384E-05	0.003597721
ZFC3H1	2.237285952	7.50363E-05	0.012456027
LUC7L	2.491726864	0.000102324	0.012980354
RALY	1.690937799	0.000109473	0.012980354
PDCD6	1.666612036	0.000297788	0.030895542
PHAX	2.908296098	0.0004174	0.038493585
MTR4	1.630218721	0.00069137	0.057383678
SNRNP70	2.031091413	0.000800001	0.060363699
HNRNPC	1.919537412	0.00102845	0.065662601
PRPF40A	1.366741504	0.00145551	0.086290926
SNRPD2	1.107723569	0.001628387	0.090104106
AGGF1	2.220149266	0.002323802	0.120547246
CASC3	1.082997907	0.006520053	0.284823385
PTCD3	1.515297002	0.007328348	0.304126433
SNRPA	2.332481199	0.009122535	0.336326781
RBM8A	1.353795952	0.009725112	0.336326781
SNRPB|SNRPN	1.424693232	0.012669024	0.413851423

Table B.7 – Continued on next page

Table B.7 – Identified ARS2 interaction partners (continued)

PROTEIN	log2 (fold change)	p-value	adjusted p-value
PRPF3	1.224829507	0.0131079	0.413851423
MRPS27	1.066498069	0.019503793	0.578148159
PPP1CC	1.210117189	0.022473688	0.621772048
CPSF2	1.146448126	0.027790083	0.63869136
SF3A1	1.285215854	0.028188845	0.63869136
EGFP- ARS2 wt - EGFP-ARS2 Δ692-743 GSG			
WDR82	2.534294659	1.78553E-06	0.001481986
NCBP3	3.448059261	5.20697E-06	0.001675443
ZC3H4	3.400909638	6.05582E-06	0.001675443
THRAP3	2.100454417	1.19078E-05	0.002470869
RALY	1.897515804	3.65205E-05	0.006062395
ZFC3H1	2.332197415	5.04741E-05	0.006982249
PHAX	3.274670495	0.000147628	0.017504487
SNRPD2	1.408418553	0.000230693	0.023934433
MTR4	1.67211756	0.000562048	0.051833353
LUC7L	2.009275546	0.000666829	0.055346829
HNRNPC	1.878201922	0.001216962	0.087132796
AGGF1	2.411878801	0.001259751	0.087132796
BCLAF1	1.122560696	0.001647468	0.105184486
PRPF6	1.009331831	0.002773502	0.164429022
SNRNP70	1.671781855	0.00338105	0.187084769
RBM8A	1.371631009	0.009009209	0.383771456
CASC3	1.02544145	0.009067173	0.383771456
SNRPE	1.579258951	0.009247505	0.383771456
CPSF3	1.311876871	0.010347505	0.408972807
PRPF40A	1.001083027	0.011308433	0.426636353
SNRPB | SNRPN	1.356688842	0.016485882	0.547331274
ALY/REF	1.015431383	0.023600326	0.653489881
SNRPA	1.932595581	0.024803708	0.653489881

BIBLIOGRAPHY

[1] N. Abovich, X. C. Liao, and M. Rosbash. "The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition". In: *Genes and Development* 8.7 (1994), pp. 843–854.

[2] M. S. Akhtar, M. Heidemann, J. R. Tietjen, D. W. Zhang, R. D. Chapman, D. Eick, and A. Z. Ansari. "TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II". In: *Molecular Cell* 34.3 (2009), pp. 387–393.

[3] T. R. Albrecht and E. J. Wagner. "snRNA 3′ end formation requires heterodimeric association of integrator subunits". In: *Molecular and Cellular Biology* 32.6 (2012), pp. 1112–1123.

[4] C. Allmang, J. Kufel, G. Chanfreau, P. Mitchell, E. Petfalski, and D. Tollervey. "Functions of the exosome in rRNA, snoRNA and snRNA synthesis". In: *EMBO Journal* 18.19 (1999), pp. 5399–5410.

[5] A. Amsterdam, R. M. Nissen, Z. Sun, E. C. Swindell, S. Farrington, and N. Hopkins. "Identification of 315 genes essential for early zebrafish development". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.35 (2004), pp. 12792–12797.

[6] P. R. Andersen, M. Domanski, M. S. Kristiansen, H. Storvall, E. Ntini, C. Verheggen, A. Schein, J. Bunkenborg, I. Poser, M. Hallais, R. Sandberg, A. Hyman, J. LaCava, M. P. Rout, J. S. Andersen, E. Bertrand, and T. H. Jensen. "The human cap-binding complex is functionally connected to the nuclear RNA exosome". In: *Nature Structural & Molecular Biology* 20.12 (2013), pp. 1367–1376.

[7] E. D. Andrulis, E. Guzman, P. Doring, J. Werner, and J. T. Lis. "High-resolution localization of Drosophila Spt5 and Spt6 at heat shock genes in vivo: roles in promoter proximal pausing and transcription elongation". In: *Genes and Development* 14.20 (2000), pp. 2635–2649.

[8] H. Ashkenazy, S. Abadi, E. Martz, O. Chay, I. Mayrose, T. Pupko, and N. Ben-Tal. "ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules". In: *Nucleic Acids Research* 44.W1 (2016), W344–350.

[9] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal. "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids". In: *Nucleic Acids Research* 38.Web Server issue (2010), W529–533.

[10] P. S. Bagga, L. P. Ford, F. Chen, and J. Wilusz. "The G-rich auxiliary downstream element has distinct sequence and position requirements and mediates efficient 3′ end pre-mRNA processing through a trans-acting factor". In: *Nucleic Acids Research* 23.9 (1995), pp. 1625–1631.

[11] D. Baillat, M. A. Hakimi, A. M. Naar, A. Shilatifard, N. Cooch, and R. Shiekhattar. "Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II". In: *Cell* 123.2 (2005), pp. 265–276.

[12] A. G. Baltz, M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, and M. Landthaler. "The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts". In: *Molecular Cell* 46.5 (2012), pp. 674–690.

[13] D. Barcaroli, L. Bongiorno-Borbone, A. Terrinoni, T. G. Hofmann, M. Rossi, R. A. Knight, A. G. Matera, G. Melino, and V. De Laurenzi. "FLASH is required for histone transcription and S-phase progression." In: *Proceedings of the National Academy of Sciences of the United States of America* 103 (40 2006), pp. 14808–14812.

[14] R. Baskaran, M. E. Dahmus, and J. Y. Wang. "Tyrosine phosphorylation of mammalian RNA polymerase II carboxyl-terminal domain". In: *Proceedings of the National Academy of Sciences of the United States of America* 90.23 (1993), pp. 11167–11171.

[15] S. Boulon, E. Basyuk, J. M. Blanchard, E. Bertrand, and C. Verheggen. "Intranuclear RNA trafficking: insights from live cell imaging". In: *Biochimie* 84.8 (2002), pp. 805–813.

[16] S. Boulon, C. Verheggen, B. E. Jady, C. Girard, C. Pescia, C. Paul, J. K. Ospina, T. Kiss, A. G. Matera, M. Bordonne, and E. Bertrand. "PHAX and CRM1 are required sequentially to transport U3 snoRNA to nucleoli". In: *Molecular Cell* 16.5 (2004), pp. 777–787.

[17] M. M. Bradford. "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding". In: *Analytical Biochemistry* 72 (1976), pp. 248–254.

[18] G. Bricogne, E. Blanc, M. Brandl, C. Flensburg, P. Keller, W. Paciorek, P. Roversi, A. Sharff, O. S. Smart, C. Vonrhein, and T. O. Womack. *BUSTER*. Cambridge, United Kingdom: Global Phasing Ltd. (2016).

[19] K. M. Brown and G. M. Gilmartin. "A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im". In: *Molecular Cell* 12.6 (2003), pp. 1467–1476.

[20] H. Bunch, X. Zheng, A. Burkholder, S. T. Dillon, S. Motola, G. Birrane, C. C. Ebmeier, S. Levine, D. Fargo, G. Hu, D. J. Taatjes, and S. K. Calderwood. "TRIM28 regulates RNA polymerase II promoter-proximal pausing and pause release." In: *Nature Structural & Molecular Biology* 21 (10 2014), pp. 876–883.

[21] A. J. C Quaresma, A. Bugai, and M. Barboric. "Cracking the control of RNA polymerase II elongation by 7SK snRNP and P-TEFb." In: *Nucleic Acids Research* 44 (16 2016), pp. 7527–7539.

[22] O. Calvo and J. L. Manley. "Strange bedfellows: polyadenylation factors at the promoter." In: *Genes and Development* 17 (11 2003), pp. 1321–1327.

[23] P. Carbon, S. Murgo, J. P. Ebel, A. Krol, G. Tebb, and L. W. Mattaj. "A common octamer motif binding protein is involved in the transcription of U6 snRNA by RNA polymerase III and U2 snRNA by RNA polymerase II". In: *Cell* 51.1 (1987), pp. 71–79.

[24] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davey, D. T. Humphreys, T. Preiss, L. M. Steinmetz, J. Krijgsveld, and M. W. Hentze. "Insights into RNA biology from an atlas of mammalian mRNA-binding proteins". In: *Cell* 149.6 (2012), pp. 1393–1406.

[25] G. Celniker, G. Nimrod, H. Ashkenazy, F. Glaser, E. Martz, I. Mayrose, T. Pupko, and N. Ben-Tal. "ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function". In: *Israel Journal of Chemistry* 53.3-4 (2013), pp. 199–206.

[26] R. S. Chambers and M. E. Dahmus. "Purification and characterization of a phosphatase from HeLa cells which dephosphorylates the C-terminal domain of RNA polymerase II". In: *Journal of Biological Chemistry* 269.42 (1994), pp. 26243–26248.

[27] C. T. Chang, G. M. Hautbergue, M. J. Walsh, N. Viphakone, T. B. van Dijk, S. Philipsen, and S. A. Wilson. "Chtop is a component of the dynamic TREX mRNA export complex". In: *EMBO Journal* 32.3 (2013), pp. 473–486.

[28] R. D. Chapman, M. Heidemann, T. K. Albert, R. Mailhammer, A. Flatley, M. Meisterernst, E. Kremmer, and D. Eick. "Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7". In: *Science* 318.5857 (2007), pp. 1780–1782.

[29] J. A. Chekanova and D. A. Belostotsky. "MicroRNAs and messenger RNA turnover." In: *Methods in molecular biology (Clifton, N.J.)* 342 (2006), pp. 73–85.

[30] F. Chen, C. C. MacDonald, and J. Wilusz. "Cleavage site determinants in the mammalian polyadenylation signal". In: *Nucleic Acids Research* 23.14 (1995), pp. 2614–2620.

[31] V. B. Chen, 3. Arendall W. B., J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson. "MolProbity: all-atom structure validation for macromolecular crystallography". In: *Acta Crystallographica. Section D, Biological Crystallography* 66.Pt 1 (2010), pp. 12–21.

[32] B. Cheng, T. Li, P. B. Rahl, T. E. Adamson, N. B. Loudas, J. Guo, K. Varzavand, J. J. Cooper, X. Hu, A. Gnatt, R. A. Young, and D. H. Price. "Functional association of Gdown1 with RNA polymerase II poised on human genes." In: *Molecular cell* 45 (1 2012), pp. 38–50.

[33] H. Cheng, K. Dufu, C. S. Lee, J. L. Hsu, A. Dias, and R. Reed. "Human mRNA export machinery recruited to the 5′ end of mRNA". In: *Cell* 127.7 (2006), pp. 1389–1400.

[34] A. Chlebowski, M. Lubas, T. H. Jensen, and A. Dziembowski. "RNA decay machines: the exosome". In: *Biochimica et Biophysica Acta* 1829.6-7 (2013), pp. 552–560.

[35] E. J. Cho, M. S. Kobor, M. Kim, J. Greenblatt, and S. Buratowski. "Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain". In: *Genes and Development* 15.24 (2001), pp. 3319–3329.

[36] G. Cooper. *The Cell: A Molecular Approach*. Sinauer Associates Inc, 2000.

[37] J. L. Corden. "Tails of RNA polymerase II". In: *Trends in Biochemical Sciences* 15.10 (1990), pp. 383–387.

[38] B. Culjkovic, I. Topisirovic, L. Skrabanek, M. Ruiz-Gutierrez, and K. L. Borden. "eIF4E is a central node of an RNA regulon that governs cellular proliferation". In: *Journal of Cell Biology* 175.3 (2006), pp. 415–426.

[39] L. Davidson, A. Kerr, and S. West. "Co-transcriptional degradation of aberrant pre-mRNA by Xrn2." In: *The EMBO journal* 31 (11 2012), pp. 2566–2578.

[40] S. M. Dias, K. F. Wilson, K. S. Rojas, A. L. Ambrosio, and R. A. Cerione. "The molecular basis for the regulation of the cap-binding complex by the importins". In: *Nature Structural & Molecular Biology* 16.9 (2009), pp. 930–937.

[41] B. Dichtl, D. Blank, M. Sadowski, W. Hubner, S. Weiser, and W. Keller. "Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination". In: *EMBO Journal* 21.15 (2002), pp. 4125–4135.

[42] S. Djebali et al. "Landscape of transcription in human cells". In: *Nature* 489.7414 (2012), pp. 101–108.

[43] Z. Dominski, J. A. Erkmann, X. Yang, R. Sanchez, and W. F. Marzluff. "A novel zinc finger protein is associated with U7 snRNP and interacts with the stem-loop binding protein in the histone pre-mRNP to stimulate 3'-end processing". In: *Genes and Development* 16.1 (2002), pp. 58–71.

[44] Z. Dominski and W. F. Marzluff. "Formation of the 3' end of histone mRNA". In: *Gene* 239.1 (1999), pp. 1–14.

[45] Z. Dominski, X. C. Yang, and W. F. Marzluff. "The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing". In: *Cell* 123.1 (2005), pp. 37–48.

[46] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton. "JPred4: a protein secondary structure prediction server." In: *Nucleic Acids Research* 43 (W1 2015), W389–394.

[47] K. Dufu, M. J. Livingstone, J. Seebacher, S. P. Gygi, S. A. Wilson, and R. Reed. "ATP is required for interactions between UAP56 and two conserved mRNA export proteins, Aly and CIP29, to assemble the TREX complex". In: *Genes and Development* 24.18 (2010), pp. 2043–2053.

[48] S. Egloff, D. O'Reilly, R. D. Chapman, A. Taylor, K. Tanzhaus, L. Pitts, D. Eick, and S. Murphy. "Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression". In: *Science* 318.5857 (2007), pp. 1777–1779.

[49] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. "Features and Development of Coot". In: *Acta Crystallographica. Section D, Biological Crystallography* 66 (2010), pp. 486–501.

[50] R. M. Esnouf. "An extensively modified version of MolScript that includes greatly enhanced coloring capabilities". In: *Journal of Molecular Graphics and Modelling* 15.2 (1997), pp. 132–134.

[51] R. M. Esnouf. "Further additions to MolScript version 1.4, including reading and contouring of electron-density maps". In: *Acta Crystallographica Section D: Biological Crystallography* 55.4 (1999), pp. 938–940.

[52] C. Fabrega, V. Shen, S. Shuman, and C. D. Lima. "Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II". In: *Molecular Cell* 11.6 (2003), pp. 1549–1561.

[53] S. Falk, K. Finogenova, M. Melko, C. Benda, S. Lykke-Andersen, T. H. Jensen, and E. Conti. "Structure of the RBM7-ZCCHC8 core of the NEXT complex reveals connections to splicing factors". In: *Nature Communications* 7 (2016), p. 13573.

[54] W. J. Feaver, J. Q. Svejstrup, N. L. Henry, and R. D. Kornberg. "Relationship of CDK-activating kinase and RNA polymerase II CTD kinase TFIIH/TFIIK". In: *Cell* 79.6 (1994), pp. 1103–1109.

[55] H. E. Feilotter, G. J. Hannon, C. J. Ruddell, and D. Beach. "Construction of an improved host strain for two hybrid screening". In: *Nucleic Acids Research* 22.8 (1994), pp. 1502–1503.

[56] S. M. Flaherty, P. Fortes, E. Izaurralde, I. W. Mattaj, and G. M. Gilmartin. "Participation of the nuclear cap binding complex in pre-mRNA 3′ processing". In: *Proceedings of the National Academy of Sciences of the United States of America* 94.22 (1997), pp. 11893–11898.

[57] M. Fornerod, M. Ohno, M. Yoshida, and I. W. Mattaj. "CRM1 is an export receptor for leucine-rich nuclear export signals". In: *Cell* 90.6 (1997), pp. 1051–1060.

[58] P. Fortes, J. Kufel, M. Fornerod, M. Polycarpou-Schwarz, D. Lafontaine, D. Tollervey, and I. W. Mattaj. "Genetic and physical interactions involving the yeast nuclear cap-binding complex". In: *Molecular and Cellular Biology* 19.10 (1999), pp. 6543–6553.

[59] K. Friend, A. F. Lovejoy, and J. A. Steitz. "U2 snRNP binds intronless histone pre-mRNAs to facilitate U7-snRNP-dependent 3′ end formation". In: *Molecular Cell* 28.2 (2007), pp. 240–252.

[60] K. Fujinaga, D. Irwin, Y. Huang, R. Taube, T. Kurosu, and B. M. Peterlin. "Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element". In: *Molecular and Cellular Biology* 24.2 (2004), pp. 787–795.

[61] M. Fukuda, S. Asano, T. Nakamura, M. Adachi, M. Yoshida, M. Yanagida, and E. Nishida. "CRM1 is responsible for intracellular transport mediated by the nuclear export signal". In: *Nature* 390.6657 (1997), pp. 308–311.

[62] J. Gabadinho et al. "MxCuBE: a synchrotron beamline control environment customized for macromolecular crystallography experiments". In: *Journal of Synchrotron Radiat.* 17.5 (2010), pp. 700–707.

[63] A. Gardini, D. Baillat, M. Cesaroni, D. Hu, J. M. Marinis, E. J. Wagner, M. A. Lazar, A. Shilatifard, and R. Shiekhattar. "Integrator regulates transcriptional initiation and pause release following activation". In: *Molecular Cell* 56.1 (2014), pp. 128–139.

[64] A. Gebhardt, M. Habjan, C. Benda, A. Meiler, D. A. Haas, M. Y. Hein, A. Mann, M. Mann, B. Habermann, and A. Pichlmair. "mRNA export through an additional cap-binding complex consisting of NCBP1 and NCBP3". In: *Nature Communications* 6 (2015), p. 8192.

[65] A. Ghosh, S. Shuman, and C. D. Lima. "Structural insights to how mammalian capping enzyme reads the CTD code". In: *Molecular Cell* 43.2 (2011), pp. 299–310.

[66] S. Giacometti, N. E. Benbahouche, M. Domanski, M. C. Robert, N. Meola, M. Lubas, J. Bukenborg, J. S. Andersen, W. M. Schulze, C. Verheggen, G. Kudla, T. H. Jensen, and E. Bertrand. "Mutually Exclusive CBC-Containing Complexes Contribute to RNA Fate". In: *Cell Reports* 18.11 (2017), pp. 2635–2650.

[67] G. M. Gilmartin. "Eukaryotic mRNA 3′ processing: a common means to different ends". In: *Genes and Development* 19.21 (2005), pp. 2517–2521.

[68] K. Glover-Cutter, S. Larochelle, B. Erickson, C. Zhang, K. Shokat, R. P. Fisher, and D. L. Bentley. "TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II". In: *Molecular and Cellular Biology* 29.20 (2009), pp. 5455–5464.

[69] G. Golling, A. Amsterdam, Z. Sun, M. Antonelli, E. Maldonado, W. Chen, S. Burgess, M. Haldi, K. Artzt, S. Farrington, S. Y. Lin, R. M. Nissen, and N. Hopkins. "Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development". In: *Nature Genetics* 31.2 (2002), pp. 135–140.

[70] D. Gorlich, R. Kraft, S. Kostka, F. Vogel, E. Hartmann, R. A. Laskey, I. W. Mattaj, and E. Izaurralde. "Importin provides a link between nuclear protein import and U snRNA export". In: *Cell* 87.1 (1996), pp. 21–32.

[71] J. J. Gruber, S. H. Olejniczak, J. Yong, G. La Rocca, G. Dreyfuss, and C. B. Thompson. "Ars2 Promotes Proper Replication-Dependent Histone mRNA 3 ′ End Formation". In: *Molecular Cell* 45.1 (2012), pp. 87–98.

[72] J. J. Gruber, D. S. Zatechka, L. R. Sabin, J. Yong, J. J. Lum, M. Kong, W. X. Zong, Z. X. Zhang, C. K. Lau, J. Rawlings, S. Cherry, J. N. Ihle, G. Dreyfuss, and C. B. Thompson. "Ars2 Links the Nuclear Cap-Binding Complex to RNA Interference and Cell Proliferation". In: *Cell* 138.2 (2009), pp. 328–339.

[73] M. Hallais, F. Pontvianne, P. R. Andersen, M. Clerici, D. Lener, N. E. Benbahouche, T. Gostan, F. Vandermoere, M. C. Robert, S. Cusack, C. Verheggen, T. H. Jensen, and E. Bertrand. "CBC-ARS2 stimulates 3 '-end maturation of multiple RNA families and favors cap-proximal processing". In: *Nature Structural & Molecular Biology* 20.12 (2013), pp. 1358–1366.

[74] J. Hamm and I. W. Mattaj. "Monomethylated cap structures facilitate RNA export from the nucleus". In: *Cell* 63.1 (1990), pp. 109–118.

[75] M. J. Hangauer, I. W. Vaughn, and M. T. McManus. "Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs". In: *PLOS Genetics* 9.6 (2013), e1003569.

[76] J. W. Harper, G. R. Adami, N. Wei, K. Keyomarsi, and S. J. Elledge. "The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases". In: *Cell* 75.4 (1993), pp. 805–816.

[77] G. M. Hautbergue, M. L. Hung, A. P. Golovanov, L. Y. Lian, and S. A. Wilson. "Mutually exclusive interactions drive handover of mRNA from export adaptors to TAP". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.13 (2008), pp. 5154–5459.

[78] G. M. Hautbergue, M. L. Hung, M. J. Walsh, A. P. Snijders, C. T. Chang, R. Jones, C. P. Ponting, M. J. Dickman, and S. A. Wilson. "UIF, a New mRNA export adaptor that works together with REF/ALY, requires FACT for recruitment to mRNA". In: *Current Biology* 19.22 (2009), pp. 1918–1924.

[79] N. Hernandez. "Formation of the 3' end of U1 snRNA is directed by a conserved sequence located downstream of the coding region". In: *EMBO Journal* 4.7 (1985), pp. 1827–1837.

[80] N. Hernandez and R. Lucito. "Elements required for transcription initiation of the human U2 snRNA gene coincide with elements required for snRNA 3' end formation". In: *EMBO Journal* 7.10 (1988), pp. 3125–3134.

[81] C. Hintermair, K. Voss, I. Forne, M. Heidemann, A. Flatley, E. Kremmer, A. Imhof, and D. Eick. "Specific threonine-4 phosphorylation and function of RNA polymerase II CTD during M phase progression". In: *Scientific Reports* 6 (2016), p. 27401.

[82] Y. Hirose and J. L. Manley. "RNA polymerase II and the integration of nuclear events." In: *Genes and Development* 14 (12 2000), pp. 1415–1429.

[83] C. K. Ho and S. Shuman. "Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme". In: *Molecular Cell* 3.3 (1999), pp. 405–411.

[84] S. N. Ho, H. D. Hunt, R. M. Horton, J. K. Pullen, and L. R. Pease. "Site-directed mutagenesis by overlap extension using the polymerase chain reaction". In: *Gene* 77.1 (1989), pp. 51–59.

[85] P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, and E. Skrzypek. "PhosphoSitePlus, 2014: mutations, PTMs and recalibrations". In: *Nucleic Acids Research* 43.Database issue (2015), pp. D512–520.

[86] N. Hosoda, Y. K. Kim, F. Lejeune, and L. E. Maquat. "CBP80 promotes interaction of Upf1 with Upf2 during nonsense-mediated mRNA decay in mammalian cells". In: *Nature Structural & Molecular Biology* 12.10 (2005), pp. 893–901.

[87] D. Hrossova, T. Sikorsky, D. Potesil, M. Bartosovic, J. Pasulka, Z. Zdrahal, R. Stefl, and S. Vanacova. "RBM7 subunit of the NEXT complex binds U-rich sequences and targets 3'-end extended forms of snRNAs". In: *Nucleic Acids Research* 43.8 (2015), pp. 4236–4248.

[88] J. P. Hsin, A. Sheth, and J. L. Manley. "RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3' end processing". In: *Science* 334.6056 (2011), pp. 683–686.

[89] J. Hu, C. S. Lutz, J. Wilusz, and B. Tian. "Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation". In: *RNA* 11.10 (2005), pp. 1485–1493.

[90] M. F. Incardona, G. P. Bourenkov, K. Levik, R. A. Pieritz, A. N. Popov, and O. Svensson. "EDNA: a framework for plugin-based applications applied to X-ray experiment online data analysis". In: *Journal of Synchrotron Radiat.* 16.Pt 6 (2009), pp. 872–879.

[91] T. Ishida and K. Kinoshita. "PrDOS: prediction of disordered protein regions from amino acid sequence". In: *Nucleic Acids Research* 35.Web Server issue (2007), W460–464.

[92] O. Isken and L. E. Maquat. "The multiple lives of NMD factors: balancing roles in gene and genome regulation". In: *Nature Reviews. Genetics* 9.9 (2008), pp. 699–712.

[93] E. Izaurralde, J. Lewis, C. Gamberi, A. Jarmolowski, C. McGuigan, and I. W. Mattaj. "A cap-binding protein complex mediating U snRNA export". In: *Nature* 376.6542 (1995), pp. 709–712.

[94] E. Izaurralde, J. Lewis, C. McGuigan, M. Jankowska, E. Darzynkiewicz, and I. W. Mattaj. "A nuclear cap binding protein complex involved in pre-mRNA splicing". In: *Cell* 78.4 (1994), pp. 657–668.

[95] E. Izaurralde, C. McGuigan, and I. W. Mattaj. "Nuclear localization of a cap-binding protein complex". In: *Cold Spring Harb Symp Quant Biol* 60 (1995), pp. 669–675.

[96] M. S. Jurica and M. J. Moore. "Pre-mRNA splicing: awash in a sea of proteins". In: *Molecular Cell* 12.1 (2003), pp. 5–14.

[97] W. Kabsch. "Xds". In: *Acta Crystallographica. Section D, Biological Crystallography* 66.Pt 2 (2010), pp. 125–132.

[98] J. Katahira, K. Strasser, A. Podtelejnikov, M. Mann, J. U. Jung, and E. Hurt. "The Mex67p-mediated nuclear mRNA export pathway is conserved from yeast to human". In: *EMBO Journal* 18.9 (1999), pp. 2593–2609.

[99] N. Kataoka, M. Ohno, I. Moda, and Y. Shimura. "Identification of the factors that interact with NCBP, an 80 kDa nuclear cap binding protein". In: *Nucleic Acids Research* 23.18 (1995), pp. 3638–3641.

[100] Y. Kerwitz, U. Kuhn, H. Lilie, A. Knoth, T. Scheuermann, H. Friedrich, E. Schwarz, and E. Wahle. "Stimulation of poly(A) polymerase through a direct interaction with the nuclear poly(A) binding protein allosterically regulated by RNA". In: *EMBO Journal* 22.14 (2003), pp. 3705–3714.

[101] H. Kim, B. Erickson, W. Luo, D. Seward, J. H. Graber, D. D. Pollock, P. C. Megee, and D. L. Bentley. "Gene-specific RNA polymerase II phosphorylation and the CTD code". In: *Nature Structural & Molecular Biology* 17.10 (2010), pp. 1279–1286.

[102] K. M. Kim, H. Cho, K. Choi, J. Kim, B. W. Kim, Y. G. Ko, S. K. Jang, and Y. K. Kim. "A new MIF4G domain-containing protein, CTIF, directs nuclear cap-binding protein CBP80/20-dependent translation". In: *Genes and Development* 23.17 (2009), pp. 2033–2045.

[103] T. Kimura, I. Hashimoto, T. Nagase, and J. Fujisawa. "CRM1-dependent, but not ARE-mediated, nuclear export of IFN-alpha1 mRNA". In: *Journal of Cell Science* 117.Pt 11 (2004), pp. 2259–2270.

[104] M. Kiriyama, Y. Kobayashi, M. Saito, F. Ishikawa, and S. Yonehara. "Interaction of FLASH with arsenite resistance protein 2 is involved in cell cycle progression at S phase". In: *Molecular and Cellular Biology* 29.17 (2009), pp. 4729–4741.

[105] S. Kitao, A. Segref, J. Kast, M. Wilm, I. W. Mattaj, and M. Ohno. "A compartmentalized phosphorylation/dephosphorylation system that regulates U snRNA export from the nucleus". In: *Molecular and Cellular Biology* 28.1 (2008), pp. 487–497.

[106] K. O. Kizer, H. P. Phatnani, Y. Shibata, H. Hall, A. L. Greenleaf, and B. D. Strahl. "A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation". In: *Molecular and Cellular Biology* 25.8 (2005), pp. 3305–3316.

[107] S. Krishnamurthy, X. He, M. Reyes-Reyes, C. Moore, and M. Hampsey. "Ssu72 Is an RNA polymerase II CTD phosphatase". In: *Molecular Cell* 14.3 (2004), pp. 387–394.

[108] E. Krissinel and K. Henrick. "Inference of macromolecular assemblies from crystalline state". In: *Journal of Molecular Biology* 372.3 (2007), pp. 774–797.

[109] U. Kuhn, M. Gundel, A. Knoth, Y. Kerwitz, S. Rudel, and E. Wahle. "Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor". In: *Journal of Biological Chemistry* 284.34 (2009), pp. 22803–22814.

[110] J. LaCava, J. Houseley, C. Saveanu, E. Petfalski, E. Thompson, A. Jacquier, and D. Tollervey. "RNA degradation by the exosome is promoted by a nuclear polyadenylation complex". In: *Cell* 121.5 (2005), pp. 713–724.

[111] A. I. Lamond and J. E. Sleeman. "Nuclear substructure and dynamics". In: *Current Biology* 13.21 (2003), R825–828.

[112] M. Landau, I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko, and N. Ben-Tal. "ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures". In: *Nucleic Acids Research* 33.Web Server issue (2005), W299–302.

[113] S. Laubinger, T. Sachsenberg, G. Zeller, W. Busch, J. U. Lohmann, G. Ratsch, and D. Weigel. "Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.25 (2008), pp. 8795–8800.

[114] B. J. Lee, A. E. Cansizoglu, K. E. Suel, T. H. Louis, Z. Zhang, and Y. M. Chook. "Rules for nuclear localization sequence recognition by karyopherin beta 2". In: *Cell* 126.3 (2006), pp. 543–558.

[115] J.-H. Lee and D. G. Skalnik. "Wdr82 is a C-terminal domain-binding protein that recruits the Setd1A Histone H3-Lys4 methyltransferase complex to transcription start sites of transcribed human genes." In: *Molecular and Cellular Biology* 28 (2 2008), pp. 609–618.

[116] N. N. Lee, V. R. Chalamcharla, F. Reyes-Turcu, S. Mehta, M. Zofall, V. Balachandran, J. Dhakshnamoorthy, N. Taneja, S. Yamanaka, M. Zhou, and S. I. Grewal. "Mtr4-like protein coordinates nuclear RNA processing for heterochromatin assembly and for telomere maintenance". In: *Cell* 155.5 (2013), pp. 1061–1074.

[117] A. Leitner, T. Walzthoeni, and R. Aebersold. "Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline." In: *Nature Protocols* 9 (1 2014), pp. 120–137.

[118] F. Lejeune, Y. Ishigaki, X. Li, and L. E. Maquat. "The exon junction complex is detected on CBP80-bound but not eIF4E-bound mRNA in mammalian cells: dynamics of mRNP remodeling". In: *EMBO Journal* 21.13 (2002), pp. 3536–3545.

[119] I. Lemm, C. Girard, A. N. Kuhn, N. J. Watkins, M. Schneider, R. Bordonne, and R. Luhrmann. "Ongoing U snRNP biogenesis is required for the integrity of Cajal bodies". In: *Molecular Biology of the Cell* 17.7 (2006), pp. 3221–3231.

[120] T. Lenasi, B. M. Peterlin, and M. Barboric. "Cap-binding protein complex links pre-mRNA capping to transcription elongation and alternative splicing through positive transcription elongation factor b (P-TEFb)". In: *Journal of Biological Chemistry* 286.26 (2011), pp. 22758–22768.

[121] A. G. Leslie. "The integration of macromolecular diffraction data". In: *Acta Crystallographica. Section D, Biological Crystallography* 62.Pt 1 (2006), pp. 48–57.

[122] J. D. Lewis and E. Izaurralde. "The role of the cap structure in RNA processing and nuclear export". In: *European Journal of Biochemistry* 247.2 (1997), pp. 461–469.

[123] W. Li, A. Cowley, M. Uludag, T. Gur, H. McWilliam, S. Squizzato, Y. M. Park, N. Buso, and R. Lopez. "The EMBL-EBI bioinformatics web and programmatic tools framework". In: *Nucleic Acids Research* 43.W1 (2015), W580–584.

[124] D. L. Lindstrom, S. L. Squazzo, N. Muster, T. A. Burckin, K. C. Wachter, C. A. Emigh, J. A. McCleery, 3. Yates J. R., and G. A. Hartzog. "Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins". In: *Molecular and Cellular Biology* 23.4 (2003), pp. 1368–1378.

[125] X. Liu, W. L. Kraus, and X. Bai. "Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways." In: *Trends in Biochemical Sciences* 40 (9 2015), pp. 516–525.

[126] Y. Liu, L. Warfield, C. Zhang, J. Luo, J. Allen, W. H. Lang, J. Ranish, K. M. Shokat, and S. Hahn. "Phosphorylation of the transcription elongation factor Spt5 by yeast Bur1 kinase stimulates recruitment of the PAF complex." In: *Molecular and cellular biology* 29 (17 2009), pp. 4852–4863.

[127] X. Lu, X. Zhu, Y. Li, M. Liu, B. Yu, Y. Wang, M. Rao, H. Yang, K. Zhou, Y. Wang, Y. Chen, M. Chen, S. Zhuang, L. F. Chen, R. Liu, and R. Chen. "Multiple P-TEFbs cooperatively regulate the release of promoter-proximally paused RNA polymerase II". In: *Nucleic Acids Research* 44.14 (2016), pp. 6853–6867.

[128] M. Lubas, M. S. Christensen, M. S. Kristiansen, M. Domanski, L. G. Falkenby, S. Lykke-Andersen, J. S. Andersen, A. Dziembowski, and T. H. Jensen. "Interaction profiling identifies the human nuclear exosome targeting complex". In: *Molecular Cell* 43.4 (2011), pp. 624–637.

[129] S. Machida, H. Y. Chen, and Y. A. Yuan. "Molecular insights into miRNA processing by Arabidopsis thaliana SERRATE". In: *Nucleic Acids Research* 39.17 (2011), pp. 7828–7836.

[130] P. Maiuri, A. Knezevich, A. De Marco, D. Mazza, A. Kula, J. G. McNally, and A. Marcello. "Fast transcription rates of RNA polymerase II in human cells." In: *EMBO Reports* 12 (12 2011), pp. 1280–1285.

[131] S. S. Mandal, C. Chu, T. Wada, H. Handa, A. J. Shatkin, and D. Reinberg. "Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.20 (2004), pp. 7572–7577.

[132] C. R. Mandel, Y. Bai, and L. Tong. "Protein factors in pre-mRNA 3'-end processing." In: *Cellular and Molecular Life Sciences* 65 (7-8 2008), pp. 1099–1122.

[133] C. R. Mandel, S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley, and L. Tong. "Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease". In: *Nature* 444.7121 (2006), pp. 953–956.

[134] F. Martin, A. Schaller, S. Eglite, D. Schumperli, and B. Muller. "The gene for histone RNA hairpin binding protein is located on human chromosome 4 and encodes a novel type of RNA binding protein". In: *EMBO Journal* 16.4 (1997), pp. 769–778.

[135] S. Masuda, R. Das, H. Cheng, E. Hurt, N. Dorman, and R. Reed. "Recruitment of the human TREX complex to mRNA during splicing". In: *Genes and Development* 19.13 (2005), pp. 1512–1517.

[136] K. Masuyama, I. Taniguchi, N. Kataoka, and M. Ohno. "RNA length defines RNA export pathway". In: *Genes and Development* 18.17 (2004), pp. 2074–2085.

[137] C. Mazza, M. Ohno, A. Segref, I. W. Mattaj, and S. Cusack. "Crystal structure of the human nuclear cap binding complex". In: *Molecular Cell* 8.2 (2001), pp. 383–396.

[138] C. Mazza, A. Segref, I. W. Mattaj, and S. Cusack. "Co-crystallization of the human nuclear cap-binding complex with a m(7) GpppG cap analogue using protein engineering". In: *Acta Crystallographica. Section D, Biological Crystallography* 58 (2002), pp. 2194–2197.

[139] C. Mazza, A. Segref, I. W. Mattaj, and S. Cusack. "Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex". In: *Embo Journal* 21.20 (2002), pp. 5548–5557.

[140] A. McCloskey, I. Taniguchi, K. Shinmyozu, and M. Ohno. "hnRNP C Tetramer Measures RNA Length to Classify RNA Polymerase II Transcripts for Export". In: *Science* 335.6076 (2012), pp. 1643–1646.

[141] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read. "Phaser crystallographic software". In: *Journal of Applied Crystallography* 40.Pt 4 (2007), pp. 658–674.

[142] S. McCracken, N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S. D. Patterson, M. Wickens, and D. L. Bentley. "The C-terminal domain of RNA polymerase II couples mRNA processing to transcription". In: *Nature* 385.6614 (1997), pp. 357–361.

[143] H. McWilliam, W. Li, M. Uludag, S. Squizzato, Y. M. Park, N. Buso, A. P. Cowley, and R. Lopez. "Analysis Tool Web Services from the EMBL-EBI". In: *Nucleic Acids Research* 41.Web Server issue (2013), W597–600.

[144] N. Meola, M. Domanski, E. Karadoulama, Y. Chen, C. Gentil, D. Pultz, K. Vitting-Seerup, S. Lykke-Andersen, J. S. Andersen, A. Sandelin, and T. H. Jensen. "Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts". In: *Molecular Cell* 64.3 (2016), pp. 520–533.

[145] C. Merz, H. Urlaub, C. L. Will, and R. Luhrmann. "Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment". In: *RNA* 13.1 (2007), pp. 116–128.

[146] P. Mitchell. "Exosome substrate targeting: the long and short of it". In: *Biochemical Society Transactions* 42.4 (2014), pp. 1129–1134.

[147] P. Mitchell, E. Petfalski, A. Shevchenko, M. Mann, and D. Tollervey. "The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'–>5' exoribonucleases". In: *Cell* 91.4 (1997), pp. 457–466.

[148] V. Mittal, B. Ma, and N. Hernandez. "SNAP(c): a core promoter factor with a built-in DNA-binding damper that is deactivated by the Oct-1 POU domain." In: *Genes and Development* 13 (14 1999), pp. 1807–1821.

[149] A. L. Mosley, S. G. Pattenden, M. Carey, S. Venkatesh, J. M. Gilmore, L. Florens, J. L. Workman, and M. P. Washburn. "Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation". In: *Molecular Cell* 34.2 (2009), pp. 168–178.

[150] S. Moteki and D. Price. "Functional coupling of capping and transcription of mRNA". In: *Molecular Cell* 10.3 (2002), pp. 599–609.

[151] A. Mourao, A. Varrot, C. D. Mackereth, S. Cusack, and M. Sattler. "Structure and RNA recognition by the snRNA and snoRNA transport factor PHAX". In: *RNA* 16.6 (2010), pp. 1205–1216.

[152] G. N. Murshudov, A. A. Vagin, and E. J. Dodson. "Refinement of macromolecular structures by the maximum-likelihood method". In: *Acta Crystallographica. Section D, Biological Crystallography* 53.Pt 3 (1997), pp. 240–255.

[153] A. Nag, K. Narsinh, and H. G. Martinson. "The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase". In: *Nature Structural & Molecular Biology* 14.7 (2007), pp. 662–669.

[154] A. Nag and J. A. Steitz. "Tri-snRNP-associated proteins interact with subunits of the TRAMP and nuclear exosome complexes, linking RNA decay and pre-mRNA splicing". In: *RNA Biology* 9.3 (2012), pp. 334–342.

[155] T. Narita, Y. Yamaguchi, K. Yano, S. Sugimoto, S. Chanarat, T. Wada, D. K. Kim, J. Hasegawa, M. Omori, N. Inukai, M. Endoh, T. Yamada, and H. Handa. "Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex". In: *Molecular and Cellular Biology* 23.6 (2003), pp. 1863–1873.

[156] T. Narita, T. M. C. Yung, J. Yamamoto, Y. Tsuboi, H. Tanabe, K. Tanaka, Y. Yamaguchi, and H. Handa. "NELF interacts with CBC and participates in 3 ' end processing of replication-dependent histone mRNAs". In: *Molecular Cell* 26.3 (2007), pp. 349–365.

[157] T. Nojima, T. Hirose, H. Kimura, and M. Hagiwara. "The interaction between cap-binding complex and RNA export factor is required for intronless mRNA export". In: *Journal of Biological Chemistry* 282.21 (2007), pp. 15645–15651.

[158] M. Ohno, A. Segref, A. Bachi, M. Wilm, and I. W. Mattaj. "PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation". In: *Cell* 101.2 (2000), pp. 187–198.

[159] B. Ossareh-Nazari, F. Bachelerie, and C. Dargemont. "Evidence for a role of CRM1 in signal-mediated nuclear protein export". In: *Science* 278.5335 (1997), pp. 141–144.

[160] C. O'Sullivan, J. Christie, M. Pienaar, J. Gambling, P. E. B. Nickerson, S. C. Alford, R. L. Chow, and P. L. Howard. "Mutagenesis of ARS2 Domains To Assess Possible Roles in Cell Cycle Progression and MicroRNA and Replication-Dependent Histone mRNA Biogenesis". In: *Molecular and Cellular Biology* 35.21 (2015), pp. 3753–3767.

[161] M. Pabis, N. Neufeld, Y. Shav-Tal, and K. M. Neugebauer. "Binding properties and dynamic localization of an alternative isoform of the cap-binding complex subunit CBP20". In: *Nucleus* 1.5 (2010), pp. 412–421.

[162] M. Pabis, N. Neufeld, M. C. Steiner, T. Bojic, Y. Shav-Tal, and K. M. Neugebauer. "The nuclear cap-binding complex interacts with the U4/U6.U5 tri-snRNP and promotes spliceosome assembly in mammalian cells". In: *RNA* 19.8 (2013), pp. 1054–1063.

[163] J. M. Pagano, H. Kwak, C. T. Waters, R. O. Sprouse, B. S. White, A. Ozer, K. Szeto, D. Shalloway, H. G. Craighead, and J. T. Lis. "Defining NELF-E RNA Binding in HIV-1 and Promoter-Proximal Pause Regions". In: *PLOS Genetics* 10.1 (2014).

[164] K. S. Paithankar and E. F. Garman. "Know your dose: RADDOSE". In: *Acta Crystallographica. Section D, Biological Crystallography* 66.Pt 4 (2010), pp. 381–388.

[165] Y. Pei and S. Shuman. "Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5". In: *Journal of Biological Chemistry* 277.22 (2002), pp. 19639–19648.

[166] B. M. Peterlin and D. H. Price. "Controlling the elongation phase of transcription with P-TEFb". In: *Molecular Cell* 23.3 (2006), pp. 297–305.

[167] F. Picard-Jean, I. Bougie, S. Shuto, and M. Bisaillon. "The immunosuppressive agent mizoribine monophosphate is an inhibitor of the human RNA capping enzyme". In: *PLOS ONE* 8.1 (2013), e54621.

[168] A. N. Popov and G. P. Bourenkov. "Choice of data-collection parameters based on statistic modelling". In: *Acta Crystallographica. Section D, Biological Crystallography* 59.Pt 7 (2003), pp. 1145–1153.

[169] D. H. Price. "P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II". In: *Molecular and Cellular Biology* 20.8 (2000), pp. 2629–2634.

[170] M. J. Prigge and D. R. Wagner. "The Arabidopsis SERRATE Gene Encodes a Zinc-Finger Protein Required for Normal Shoot Development". In: *Plant Cell* 13.6 (2001), pp. 1263–1279.

[171] N. Proudfoot. "New perspectives on connecting messenger RNA 3′ end formation to transcription." In: *Current Opinion in Cell Biology* 16 (3 2004), pp. 272–278.

[172] N. J. Proudfoot. "Ending the message: poly(A) signals then and now." In: *Genes and Development* 25 (17 2011), pp. 1770–1782.

[173] N. J. Proudfoot, A. Furger, and M. J. Dye. "Integrating mRNA processing with transcription." In: *Cell* 108 (4 2002), pp. 501–512.

[174] J. N. Rao, L. Neumann, S. Wenzel, K. Schweimer, P. Rosch, and B. M. Wohrl. "Structural studies on the RNA-recognition motif of NELF E, a cellular negative transcription elongation factor involved in the regulation of HIV transcription". In: *Biochemical Journal* 400.3 (2006), pp. 449–456.

[175] R. J. Read, P. D. Adams, 3. Arendall W. B., A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lutteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend, and P. H. Zwart. "A new generation of crystallographic validation tools for the protein data bank". In: *Structure* 19.10 (2011), pp. 1395–1412.

[176] J. Reikofski and B. Y. Tao. "Polymerase chain reaction (PCR) techniques for site-directed mutagenesis". In: *Biotechnology Advances* 10.4 (1992), pp. 535–547.

[177] D. B. Renner, Y. Yamaguchi, T. Wada, H. Handa, and D. H. Price. "A highly purified RNA polymerase II elongation control system". In: *Journal of Biological Chemistry* 276.45 (2001), pp. 42601–42609.

[178] M. Rienzo and A. Casamassimi. "Integrator complex and transcription regulation: Recent findings and pathophysiology". In: *Biochimica et Biophysica Acta* 1859.10 (2016), pp. 1269–1280.

[179] X. Robert and P. Gouet. "Deciphering key features in protein structures with the new ENDscript server". In: *Nucleic Acids Research* 42.Web Server issue (2014), W320–324.

[180] T. G. Rossman and Z. Wang. "Expression cloning for arsenite-resistance resulted in isolation of tumor-suppressor fau cDNA: possible involvement of the ubiquitin system in arsenic carcinogenesis". In: *Carcinogenesis* 20.2 (1999), pp. 311–316.

[181] K. Ryan, O. Calvo, and J. L. Manley. "Evidence that polyadenylation factor CPSF-73 is the mRNA 3′ processing endonuclease". In: *RNA* 10.4 (2004), pp. 565–573.

[182] K. Ryan, K. G. Murthy, S. Kaneko, and J. L. Manley. "Requirements of the RNA polymerase II C-terminal domain for reconstituting pre-mRNA 3′ cleavage". In: *Molecular and Cellular Biology* 22.6 (2002), pp. 1684–1692.

[183] I. Sabath, A. Skrajna, X. C. Yang, M. Dadlez, W. F. Marzluff, and Z. Dominski. "3'-End processing of histone pre-mRNAs in Drosophila: U7 snRNP is associated with FLASH and polyadenylation factors". In: *RNA* 19.12 (2013), pp. 1726–1744.

[184] L. R. Sabin, R. Zhou, J. J. Gruber, N. Lukinova, S. Bambina, A. Berman, C. K. Lau, C. B. Thompson, and S. Cherry. "Ars2 Regulates Both miRNA- and siRNA-Dependent Silencing and Suppresses RNA Virus Infection in Drosophila". In: *Cell* 138.2 (2009), pp. 340–351.

[185] A. B. Sachs, P. Sarnow, and M. W. Hentze. "Starting at the beginning, middle, and end: translation initiation in eukaryotes." In: *Cell* 89 (6 1997), pp. 831–838.

[186] J. Sambrook, E. Fritsch, and T. Maniatis. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 1989.

[187] H. Santos-Rosa, H. Moreno, G. Simos, A. Segref, B. Fahrenkrog, N. Pante, and E. Hurt. "Nuclear mRNA export requires complex formation between Mex67p and Mtr2p at the nuclear pores". In: *Molecular and Cellular Biology* 18.11 (1998), pp. 6826–6838.

[188] A. Segref, I. W. Mattaj, and M. Ohno. "The evolutionarily conserved region of the U snRNA export mediator PHAX is a novel RNA-binding domain that is essential for U snRNA export". In: *RNA* 7.3 (2001), pp. 351–360.

[189] A. Segref, K. Sharma, V. Doye, A. Hellwig, J. Huber, R. Luhrmann, and E. Hurt. "Mex67p, a novel factor for nuclear mRNA export, binds to both poly(A)+ RNA and nuclear pores". In: *EMBO Journal* 16.11 (1997), pp. 3256–3271.

[190] A. J. Shatkin and J. L. Manley. "The ends of the affair: capping and polyadenylation". In: *Nature Structural & Molecular Biology* 7.10 (2000), pp. 838–842.

[191] E. Y. Shim, A. K. Walker, Y. Shi, and T. K. Blackwell. "CDK-9/cyclin T (P-TEFb) is required in two postinitiation pathways for transcription in the C. elegans embryo". In: *Genes and Development* 16.16 (2002), pp. 2135–2146.

[192] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". In: *Molecular Systems Biology* 7 (2011), p. 539.

[193] J. R. Skaar, A. L. Ferris, X. Wu, A. Saraf, K. K. Khanna, L. Florens, M. P. Washburn, S. H. Hughes, and M. Pagano. "The Integrator complex controls the termination of transcription at diverse classes of gene targets". In: *Cell Research* 25.3 (2015), pp. 288–305.

[194] A. Skrajna, X. C. Yang, K. Bucholc, J. Zhang, T. M. Hall, M. Dadlez, W. F. Marzluff, and Z. Dominski. "U7 snRNP is recruited to histone pre-mRNA in a FLASH-dependent manner by two separate regions of the Stem-Loop Binding Protein". In: *RNA* (2017).

[195] K. Stade, C. S. Ford, C. Guthrie, and K. Weis. "Exportin 1 (Crm1p) is an essential nuclear export factor". In: *Cell* 90.6 (1997), pp. 1041–1050.

[196] B. Stadelmayer, G. Micas, A. Gamot, P. Martin, N. Malirat, S. Koval, R. Raffel, B. Sobhian, D. Severac, S. Rialle, H. Parrinello, O. Cuvier, and M. Benkirane. "Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes". In: *Nature Communications* 5 (2014).

[197] K. Strasser and E. Hurt. "Yra1p, a conserved nuclear RNA-binding protein, interacts directly with Mex67p and is required for mRNA export". In: *EMBO Journal* 19.3 (2000), pp. 410–420.

[198] K. Strasser, S. Masuda, P. Mason, J. Pfannstiel, M. Oppizzi, S. Rodriguez-Navarro, A. G. Rondon, A. Aguilera, K. Struhl, R. Reed, and E. Hurt. "TREX is a conserved complex coupling transcription with messenger RNA export". In: *Nature* 417.6886 (2002), pp. 304–308.

[199] F. Stutz, A. Bachi, T. Doerks, I. C. Braun, B. Seraphin, M. Wilm, P. Bork, and E. Izaurralde. "REF, an evolutionary conserved family of hnRNP-like proteins, interacts with TAP/Mex67p and participates in mRNA nuclear export". In: *RNA* 6.4 (2000), pp. 638–650.

[200] I. Taniguchi and M. Ohno. "ATP-dependent recruitment of export factor Aly/REF onto intronless mRNAs by RNA helicase UAP56". In: *Molecular and Cellular Biology* 28.2 (2008), pp. 601–608.

[201] H. Uemura and Y. Jigami. "GCR3 encodes an acidic protein that is required for expression of glycolytic genes in Saccharomyces cerevisiae". In: *Journal of Bacteriology* 174.17 (1992), pp. 5526–5532.

[202] S. Vanacova, J. Wolf, G. Martin, D. Blank, S. Dettwiler, A. Friedlein, H. Langen, G. Keith, and W. Keller. "A new yeast poly(A) polymerase complex involved in RNA quality control". In: *PLOS Biology* 3.6 (2005), e189.

[203] K. Venkataraman, K. M. Brown, and G. M. Gilmartin. "Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition". In: *Genes and Development* 19.11 (2005), pp. 1315–1327.

[204] P. Vinciguerra and F. Stutz. "mRNA export: an assembly line from genes to nuclear pores." In: *Current Opinion in Cell Biology* 16 (3 2004), pp. 285–292.

[205] N. Visa, E. Izaurralde, J. Ferreira, B. Daneholt, and I. W. Mattaj. "A nuclear cap-binding complex binds Balbiani ring pre-mRNA cotranscriptionally and accompanies the ribonucleoprotein particle during nuclear export". In: *Journal of Cell Biology* 133.1 (1996), pp. 5–14.

[206] S. M. Vos, D. Pollmann, L. Caizzi, K. B. Hofmann, P. Rombaut, T. Zimniak, F. Herzog, and P. Cramer. "Architecture and RNA binding of the human negative elongation factor". In: *Elife* 5 (2016).

[207] T. Wada, T. Takagi, Y. Yamaguchi, A. Ferdous, T. Imai, S. Hirose, S. Sugimoto, K. Yano, G. A. Hartzog, F. Winston, S. Buratowski, and H. Handa. "DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs." In: *Genes and Development* 12 (3 1998), pp. 343–356.

[208] E. Wahle. "A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation". In: *Cell* 66.4 (1991), pp. 759–768.

[209] E. Wahle. "Poly(A) tail length control is caused by termination of processive synthesis". In: *Journal of Biological Chemistry* 270.6 (1995), pp. 2800–2808.

[210] A. C. Wallace, R. A. Laskowski, and J. M. Thornton. "LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions." In: *Protein Engineering* 8 (2 1995), pp. 127–134.

[211] T. Walzthoeni, M. Claassen, A. Leitner, F. Herzog, S. Bohn, F. Forster, M. Beck, and R. Aebersold. "False discovery rate estimation for cross-linked peptides identified by mass spectrometry". In: *Nature Methods* 9.9 (2012), pp. 901–903.

[212] Z. F. Wang, M. L. Whitfield, 3. Ingledue T. C., Z. Dominski, and W. F. Marzluff. "The protein that binds the 3' end of histone mRNA: a novel RNA-binding protein required for histone pre-mRNA processing". In: *Genes and Development* 10.23 (1996), pp. 3028–3040.

[213] N. J. Watkins, I. Lemm, and R. Luhrmann. "Involvement of nuclear import and export factors in U8 box C/D snoRNP biogenesis". In: *Molecular and Cellular Biology* 27.20 (2007), pp. 7018–7027.

[214] Y. Wen and A. J. Shatkin. "Transcription elongation factor hSPT5 stimulates mRNA capping". In: *Genes and Development* 13.14 (1999), pp. 1774–1779.

[215] J. W. Werner-Allen, C. J. Lee, P. Liu, N. I. Nicely, S. Wang, A. L. Greenleaf, and P. Zhou. "cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72". In: *Journal of Biological Chemistry* 286.7 (2011), pp. 5717–5726.

[216] A. J. Westermann, S. A. Gorski, and J. Vogel. "Dual RNA-seq of pathogen and host". In: *Nature Reviews. Microbiology* 10.9 (2012), pp. 618–630.

[217] M. Wickens, P. Anderson, and R. J. Jackson. "Life and death in the cytoplasm: messages from the 3' end". In: *Current Opinion in Genetics and Development* 7.2 (1997), pp. 220–232.

[218] C. L. Will and R. Luhrmann. "Spliceosome structure and function". In: *Cold Spring Harb Perspect Biol* 3.7 (2011).

[219] M. D. Wilson, D. Wang, R. Wagner, H. Breyssens, M. Gertsenstein, C. Lobe, X. Lu, A. Nagy, R. D. Burke, B. F. Koop, and P. L. Howard. "ARS2 is a conserved eukaryotic gene essential for early mammalian development". In: *Molecular and Cellular Biology* 28.5 (2008), pp. 1503–1514.

[220] C. J. Wilusz, M. Wormington, and S. W. Peltz. "The cap-to-tail guide to mRNA turnover". In: *Nature Reviews. Molecular Cell Biology* 2.4 (2001), pp. 237–246.

[221] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, and K. S. Wilson. "Overview of the CCP4 suite and current developments". In: *Acta Crystallographica. Section D, Biological Crystallography* 67.Pt 4 (2011), pp. 235–242.

[222] R. Worch, A. Niedzwiecka, J. Stepinski, C. Mazza, M. Jankowska-Anyszka, E. Darzynkiewicz, S. Cusack, and R. Stolarski. "Specificity of recognition of mRNA 5′ cap by human nuclear cap-binding complex." In: *RNA* 11 (9 2005), pp. 1355–1363.

[223] M. Wu, P. F. Wang, J. S. Lee, S. Martin-Brown, L. Florens, M. Washburn, and A. Shilatifard. "Molecular regulation of H3K4 trimethylation by Wdr82, a component of human Set1/COMPASS." In: *Molecular and Cellular Biology* 28 (24 2008), pp. 7337–7344.

[224] F. Wyers, M. Rougemaille, G. Badis, J. C. Rousselle, M. E. Dufour, J. Boulay, B. Regnault, F. Devaux, A. Namane, B. Seraphin, D. Libri, and A. Jacquier. "Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase". In: *Cell* 121.5 (2005), pp. 725–737.

[225] K. Xiang, T. Nagaike, S. Xiang, T. Kilic, M. M. Beh, J. L. Manley, and L. Tong. "Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex". In: *Nature* 467.7316 (2010), pp. 729–733.

[226] T. Yamada, Y. Yamaguchi, N. Inukai, S. Okamoto, T. Mura, and H. Handa. "P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation". In: *Molecular Cell* 21.2 (2006), pp. 227–237.

[227] Y. Yamaguchi, T. Takagi, T. Wada, K. Yano, A. Furuya, S. Sugimoto, J. Hasegawa, and H. Handa. "NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation". In: *Cell* 97.1 (1999), pp. 41–51.

[228] J. Yamamoto, Y. Hagiwara, K. Chiba, T. Isobe, T. Narita, H. Handa, and Y. Yamaguchi. "DSIF and NELF interact with Integrator to specify the correct post-transcriptional fate of snRNA genes". In: *Nature Communications* 5 (2014), p. 4263.

[229] X. C. Yang, B. D. Burch, Y. Yan, W. F. Marzluff, and Z. Dominski. "FLASH, a proapoptotic protein involved in activation of caspase-8, is essential for 3′ end processing of histone pre-mRNAs". In: *Molecular Cell* 36.2 (2009), pp. 267–278.

[230] C. Y. Yuo, J. Ares M., and A. M. Weiner. "Sequences required for 3′ end formation of human U2 small nuclear RNA". In: *Cell* 42.1 (1985), pp. 193–202.

[231] U. Zander, M. Cianci, N. Foos, C. S. Silva, L. Mazzei, C. Zubieta, A. de Maria, and M. H. Nanao. "Merging of synchrotron serial crystallographic data by a genetic algorithm". In: *Acta Crystallographica. Section D, Biological Crystallography* 72.Pt 9 (2016), pp. 1026–1035.

[232] Y. Zhang, M. Zhang, and Y. Zhang. "Crystal structure of Ssu72, an essential eukaryotic phosphatase specific for the C-terminal domain of RNA polymerase II, in complex with a transition state analogue". In: *Biochemical Journal* 434.3 (2011), pp. 435–444.

[233] Q. Zhou, T. Li, and D. H. Price. "RNA polymerase II elongation control." In: *Annual Review of Biochemistry* 81 (2012), pp. 119–143.

[234] D. A. R. Zorio and D. L. Bentley. "The link between mRNA processing and transcription: communication works both ways." In: *Experimental cell research* 296 (1 2004), pp. 91–97.

ACKNOWLEDGMENTS

First of all I want to thank my PhD supervisor Dr. Stephen Cusack for the opportunity to perform my PhD thesis within his lab and on this challenging and highly interesting project. I am grateful for his guidance, support and unbreakable believe in obtaining well diffracting crystals.

Next, many thanks to all who supported me during my time in Grenoble and at EMBL and made the last years to an unforgettable memory. Special thanks go to Emiko and Alexander for creating such a great working atmosphere and all their scientific and non-scientific advises; to Lahari, Petra, Erika, Francesco and Audrey for sharing their past four years with me and to all the other great colleagues I met and friends I made.

For the great services at the facilities I would like to thank especially Joanna Kirkpatrick, Mandy Rettel and Frank Stein (Proteomics Core Facility EMBL Heidelberg), Luca Signor (Mass Spectrometry Platform IBS Grenoble), Malte Paulsen (Flow Cytometry Facility EMBL Heidelberg) and Marc Jamin and Caroline Mas (PSB Biophysical Platform Grenoble). An extra big thanks goes to the great HTX Lab at EMBL Grenoble and all their present and past members for processing all my crystallisation requests.

Furthermore, I want to thank Matthew Bowler for providing access to the MASSIF beamline, which allowed me to measure plenty of different crystals, Max Nano for his help with the phasing of the ARS2 data sets and Thomas Bock (Martin Beck group) for sharing the cross-linking protocol and his advises.

Thanks go to our collaborator Edouard Bertrand (CNRS, Montpellier) for all the discussions regarding the CBC complexes and for allowing me to join his lab to perform some experiments. I also thank his group members for welcoming me in Montpellier - in particular Marie-Cecile, who took care of me during my stay, prepared everything for the experiments and continued them afterwards.

I would also like to acknowledge my committee members: Prof. Dr. Iain Mattaj, Prof. Dr. Georg Stoecklin and Dr. Christiane Berger-Schaffitzel for their constant scientific feedback.

I further want to thank Florian, Valentina and Stephen for their constructive feedback for this thesis and my father Arno for all his help with LyX.

Lastly, thanks to my parents, Manuela and Arno, for all the support, love and security they are giving me.

DECLARATION

I hereby declare that I have authored the doctoral thesis "Mutually Exclusive CBC and CBC-ARS2 Containing Complexes Coordinate the Fate of RNA Polymerase II Transcripts" independently, that I have not used other than the declared sources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Grenoble, June 2017

Wiebke Manuela Schulze